Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis

Brooke Levis,1 Andrea Benedetti,2 Brett D Thombs,1 on behalf of the DEPRESsion Screening Data (DEPRESsion) Collaboration

ABSTRACT

OBJECTIVE
To determine the accuracy of the Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression.

DESIGN
Individual participant data meta-analysis.

DATA SOURCES
Medline, Medline In-Process and Other Non-Indexed Citations, PsycINFO, and Web of Science (January 2000-February 2015).

INCLUSION CRITERIA
Eligible studies compared PHQ-9 scores with major depression diagnoses from validated diagnostic interviews. Primary study data and study level data extracted from primary reports were synthesized. For PHQ-9 cut-off scores 5-15, bivariate random effects meta-analysis was used to estimate pooled sensitivity and specificity, separately, among studies that used semistructured diagnostic interviews, which are designed for administration by clinicians; fully structured interviews, which are designed for lay administration; and the Mini International Neuropsychiatric (MINI) diagnostic interviews, a brief fully structured interview. Sensitivity and specificity were examined among participant subgroups and, separately, using meta-regression, considering all subgroup variables in a single model.

RESULTS
Data were obtained for 58 of 72 eligible studies (total n=17 357; major depression cases n=2312). Combined sensitivity and specificity was maximized at a cut-off score of 10 or above among studies using a semistructured interview (29 studies, 6725 participants; sensitivity 0.88, 95% confidence interval 0.83 to 0.92; specificity 0.85, 0.82 to 0.88). Across cut-off scores 5-15, sensitivity with semistructured interviews was 5-22% higher than for fully structured interviews (MINI excluded; 14 studies, 7680 participants) and 2-15% higher than for the MINI (15 studies, 2952 participants). Specificity was similar across diagnostic interviews. The PHQ-9 seems to be similarly sensitive but may be less specific for younger patients than for older patients; a cut-off score of 10 or above can be used regardless of age..

CONCLUSIONS
PHQ-9 sensitivity compared with semistructured diagnostic interviews was greater than in previous conventional meta-analyses that combined reference standards. A cut-off score of 10 or above maximized combined sensitivity and specificity overall and for subgroups.

WHAT IS ALREADY KNOWN ON THIS TOPIC
The Patient Health Questionnaire-9 (PHQ-9) is the most commonly used tool for screening for depression in primary care.

Previous meta-analyses on diagnostic test accuracy of PHQ-9 have had limitations including selective cut-off reporting in primary studies and inability to assess differences across patient subgroups.

They also did not exclude participants already diagnosed as having or being treated for depression, who would not be screened in practice.

WHAT THIS STUDY ADDS
Diagnostic accuracy of PHQ-9 compared with diagnoses made by semistructured diagnostic interviews is greater than when compared with diagnoses made by other reference standards.

Diagnostic accuracy of PHQ-9 does not differ substantively across participant subgroups except for age, where it may be more specific among older patients.

The standard cut-off score of 10 or greater maximizes combined sensitivity and specificity overall and for subgroups.

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

Introduction
Screening for depression refers to the use of a depression screening questionnaire to identify patients who may have depression but have not been identified. When screening programs are recommended, clinicians are advised to administer a depression symptom questionnaire and to use a pre-identified cut-off threshold to classify patients as having positive or negative screening results. Those with positive screening results can then be evaluated to determine whether they have depression and, if appropriate, should be offered treatment.1 2

The Patient Health Questionnaire-9 (PHQ-9) is a nine item questionnaire designed to screen for depression in primary care and other medical settings.3-7 The standard cut-off score for screening to identify possible major depression is 10 or above,3-7 which was established in the first study on the PHQ-9 (total n=580, major depression n=41).3-5

A conventional PHQ-9 meta-analysis from 2015 (36 studies, 21 292 participants) evaluated sensitivity and specificity for cut-off scores 7-15 by combining accuracy results for each cut-off score that were published in included primary studies.8 Pooled
sensitivity for the standard cut-off score of 10 was 0.78 (95% confidence interval 0.70 to 0.84), and pooled specificity was 0.87 (0.84 to 0.90). Incomplete reporting of results from cut-off scores other than 10 in the primary studies that were included, however, resulted in cut-off score ranges in which sensitivity implausibly increased as cut-off scores increased. This suggested possible selective cut-off reporting in some primary studies to maximize accuracy. Additional limitations included the inability to assess differences across patient subgroups, as subgroup results were not reported in primary studies; the inability to exclude participants already diagnosed as having or being treated for depression, who would not be screened in practice but were included in many primary studies; and the combining of accuracy estimates without differentiating between reference standards. Semistructured diagnostic interviews (for example, Structured Clinical Interview for DSM Disorders (SCID)) are intended to be employed trained clinical interviewers; they are administered by lay interviewers to reduce the cost of employing trained clinical interviewers; they are intended to achieve a high level of standardization but may sacrifice accuracy. The Mini International Neuropsychiatric Interview (MINI) is fully structured but was designed for very rapid administration and described by its authors as being overinclusive as a result. In a recent analysis, controlling for depressive symptom scores, we found that the MINI classified approximately twice as many participants as having major depression as other fully structured interviews. Compared with semistructured interviews, fully structured interviews (MINI excluded) classified more patients with low symptom scores but fewer patients with high symptom scores as having major depression.

Individual participant data meta-analysis involves a standard systematic review, then synthesis of participant level data from primary studies rather than summary results from study reports. Advantages include the ability to do subgroup analyses not reported in primary studies, the ability to report results from all relevant cut-off scores from all included studies, and the ability to exclude participants already diagnosed as having or treated for depression who would not be screened in practice. The objectives of this study were to use individual participant data meta-analysis to evaluate the diagnostic accuracy of the PHQ-9 screening tool among studies using semistructured, fully structured (MINI excluded), and MINI diagnostic interviews as reference standards, separately, with priority given to semistructured interview results; among participants not diagnosed as having or receiving treatment for a mental health problem; and among participant subgroups based on age, sex, country human development index, and recruitment setting.

Methods
This individual participant data meta-analysis was registered in PROSPERO (CRD42014010673), a protocol was published, and results were reported following PRISMA-DTA and PRISMA-IPD reporting guidelines. A medical librarian searched Medline, Embase, In-Process and Other Non-Indexed Citations via Ovid, PsycINFO, and Web of Science (January 1, 2000-February 7, 2015) on February 7, 2015, using a peer reviewed search strategy (supplementary methods A). The search was limited to 2000 forward because the PHQ-9 was published in 2001. We also reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada) for storing and tracking of search results.

Identification of eligible studies
Datasets from articles in any language were eligible for inclusion if they included diagnostic classification for current major depressive disorder or major depressive episode on the basis of a validated semistructured or fully structured interview conducted within two weeks of PHQ-9 administration among participants aged 18 years or over who were not recruited from youth or psychiatric settings or because they were identified as having symptoms of depression. We required the diagnostic interviews and PHQ-9 to be administered within two weeks of each other because the Diagnostic and Statistical Manual of Mental Disorders (DSM) and international classification of diseases (ICD) diagnostic criteria for major depression specify that symptoms must have been present in the previous two weeks. We excluded patients from psychiatric settings and those already identified as having symptoms of depression because screening is done to identify previously unrecognized cases.

Datasets in which not all participants were eligible were included if primary data allowed selection of eligible participants. For defining major depression, we considered major depressive disorder or major depressive episode based on the DSM or ICD criteria. If more than one was reported, we prioritized major depressive episode over major depressive disorder, as screening would attempt to detect depressive episodes and further interview would determine whether the episode was related to major depressive disorder or bipolar disorder, and DSM over ICD. Across all studies, there were 23 discordant diagnoses depending on classification prioritization (0.1% of participants). Two investigators independently reviewed titles and abstracts for eligibility. If either deemed a study potentially eligible, two investigators did full text review independently, with disagreements resolved by consensus, consulting a third investigator when
Data extraction, contribution, and synthesis

We invited authors of eligible datasets to contribute de-identified primary data. Two investigators independently extracted country, recruitment setting (non-medical, primary care, inpatient specialty, outpatient specialty), and diagnostic interview from published reports, with disagreements resolved by consensus. We categorized countries as “very high,” “high,” or “low-medium” development on the basis of the United Nations’ human development index.26 Participant level data included age, sex, major depression status, current mental health diagnosis or treatment, and PHQ-9 scores. In two primary studies, multiple recruitment settings were included, so recruitment setting was coded at the participant level. When datasets included statistical weights to reflect sampling procedures, we used the weights provided. For studies in which sampling procedures merited weighting but the original study did not weight, we constructed weights by using inverse selection probabilities. Weighting occurred, for instance, when a diagnostic interview was administered to all participants with positive screens and a random subset of participants with negative screens.

We converted individual participant data to a standard format and synthesized them into a single dataset with study level data. We compared published participant characteristics and diagnostic accuracy results with results from raw datasets and resolved any discrepancies in consultation with the original investigators.

Two investigators assessed risk of bias of included studies independently, on the basis of the primary publications, using the Quality Assessment of Diagnostic Accuracy Studies-2 tool (supplementary methods B).27 Discrepancies were resolved by consensus.

Statistical analyses

We did three main sets of analyses. Firstly, we estimated sensitivity and specificity across PHQ-9 cut-off scores 5-15 for studies with semistructured (SCID,23 Schedules for Clinical Assessment in Neuropsychiatry,28 Depression Interview and Structured Hamilton30), fully structured (MINI excluded; CIDI,14 Clinical Interview Schedule-Revised,30 Diagnostic Interview Schedule31), and MINI19 20 reference standards, separately. Secondly, for each reference standard category, we estimated sensitivity and specificity across PHQ-9 cut-off scores for all participants from primary studies, as has been done in existing conventional meta-analyses and, separately, among only participants who could be confirmed as not diagnosed as having or receiving treatment for a mental health problem at the time of assessment. We did this because existing conventional meta-analyses have all been based on primary studies that generally do not exclude patients already diagnosed as having or receiving treatment for a mental health problem. As screening is done to identify previously unrecognized cases, those patients would not be screened in practice, and their inclusion in diagnostic accuracy studies could bias results.10 11

Thirdly, for each reference standard category, we estimated and compared sensitivity and specificity across PHQ-9 cut-off scores among subgroups based on age (<60 v ≥60 years), sex, country human development index (very high, high, low-medium), and recruitment setting (non-medical, primary, inpatient specialty, outpatient specialty). Among studies that used the MINI, we combined inpatient and outpatient specialty care settings, as only one study included inpatient participants. In each subgroup analysis, we excluded primary studies with no major depression cases, as this did not allow application of the bivariate random effects model. This resulted in a maximum of 15 participants excluded from any subgroup analysis.

For each meta-analysis, for cut-off scores 5-15 separately, bivariate random effects models were fitted via Gauss-Hermite adaptive quadrature.32 This two stage meta-analytic approach models sensitivity and specificity simultaneously, accounting for the inherent correlation between them and for precision of estimates within studies. For each analysis, this model provided estimates of pooled sensitivity and specificity.

To compare results across reference standards and other subgroups, we constructed empirical receiver operating characteristic curves for each group based on the pooled sensitivity and specificity estimates and calculated areas under the curve. We estimated differences in sensitivity and specificity between subgroups at each cut-off score by constructing confidence intervals for differences via the cluster bootstrap approach.33 34 This resulted in a maximum of 15 participants excluded from any subgroup analysis. For each comparison, we ran 1000 iterations of the bootstrap. We removed iterations that did not produce difference estimates for cut-off scores 5-15 before determining confidence intervals and noted the number of iterations removed.

In addition to categorical subgroup analyses, we compared sensitivity and specificity across the different reference standards by doing one stage meta-regressions with interactions between reference standard category (reference category=semistructured interviews) and accuracy coefficients (logit(sensitivity)) and logit(specificity)), and we compared results with those seen in the original two stage bivariate random effects meta-analytic models. Additionally, within each reference standard category, we did one stage meta-regressions in which we interacted all subgrouping variables (age (measured continuously), sex (reference category=women), country human development index (reference category=very high), and participant recruitment setting (reference category=primary care)) with logit(sensitivity) and logit(specificity). Similarly to our main subgroup analyses, we again determined which significant interactions replicated across all three reference standard categories. For subgrouping variables that were significantly associated with sensitivity or specificity coefficients for all three reference standard
categories for all or most cut-off scores in the main one stage meta-regression, we did additional one stage meta-regression to produce accuracy estimates for the subgroups of interest, and we compared these results with those seen in the original two stage bivariate random effects meta-analytic models. Although age was included as a continuous variable in the main meta-regression, we again dichotomized it (<60 v ≥60 years) to estimate accuracy and for comparison with the bivariate model results.

To investigate heterogeneity, we generated forest plots of sensitivities and specificities for cut-off score 10 for each study, first for all studies in each reference standard category and then separately across participant subgroups within each reference standard category. We quantified cut-off score 10 heterogeneity overall and across subgroups by reporting estimated variances of the random effects for sensitivity and specificity ($\tau^2$) and 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. We used a complete case analysis, as complete data for all subgrouping variables were available for 17,357 participants (98% of eligible participants in the database).

To estimate positive and negative predictive values using cut-off score 10 for different values of prevalence of major depression, we generated nomograms for each reference standard category by applying the cut-off 10 sensitivity and specificity estimates from the meta-analysis to hypothetical major depression prevalence values of 5-25%.

In sensitivity analyses, for each reference standard category, we compared accuracy results across subgroups based on Quality Assessment of Diagnostic Accuracy Studies-2 items for all items with at least 100 cases of major depression among participants categorized as having “low” risk of bias and among participants with “high” or “unclear” risk of bias.

We did not do sensitivity analyses that combined accuracy results from individual participant data meta-analysis with published results from studies that did not contribute individual participant data, only two studies with a semistructured reference standard (total n=173, major depression n=29), one study with a fully structured reference standard (total n=730, major depression n=32), and one study using the MINI (total n=172, major depression n=33) published accuracy results eligible for this individual participant data meta-analysis. The other studies had eligible datasets but did not publish eligible diagnostic accuracy results (supplementary table A).

For all analyses, we used R (R version R 3.4.1 and R Studio version 1.0.143) using the glmer function within the lme4 package, which uses one quadrature point. The only substantive deviations from our initial protocol were that we stratified accuracy results by reference standard category and did not do sensitivity analyses that combined accuracy results from individual participant data meta-analysis with published results from studies that did not contribute individual participant data.

### Patient and public involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for design or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to study participants or the relevant patient community.

### Results

#### Search results and inclusion of primary datasets

Of 5248 unique titles and abstracts identified from the database search, 5039 were excluded after review of titles and abstracts and 113 after full text review, leaving 96 eligible articles with data from 69 unique participant samples, of which 55 (80%) contributed datasets (supplementary figure A). Reasons for exclusion for the 113 articles excluded at full text level are given in supplementary table A. In addition, authors of included studies contributed data from three unpublished studies, for a total of 58 datasets (total n=17,357, major depression n=2312 (13%)). Characteristics of included studies and eligible studies that did not provide datasets are shown in supplementary table B. Excluding the three

| Table 1 | Participant data by diagnostic interview |
|-----------------|------------------|-----------------|------------------|
| **Diagnostic interview** | **No of studies** | **No of participants** | **No (%) with major depression** |
| SCID | 26 | 4733 | 785 (17) |
| DIS | 1 | 100 | 9 (9) |
| DISH | 1 | 1006 | 221 (22) |
| CIS-R | 2 | 402 | 64 (16) |
| MINI | 15 | 2952 | 549 (19) |
| Total | 58 | 17,357 | 2312 (13) |

**CIDI=**Composite International Diagnostic Interview; **CIS-R=**Clinical Interview Schedule- Revised; **DIS=**Diagnostic Interview Schedule; **DISH=**Depression Interview and Structured Hamilton; **MINI=**Mini International Neuropsychiatric Interview; **SCAN=**Schedules for Clinical Assessment in Neuropsychiatry; **SCID=**Structured Clinical Interview for DSM Disorders.
unpublished studies, of 21 171 participants in 69 eligible published studies, 16 956 (80%) participants from 55 included published studies were included.

Of 58 included studies, 29 used semistructured reference standards, 14 used fully structured reference standards, and 15 used the MINI (table 1). The SCID was the most common semistructured interview (26 studies, total n=4733), and the CIDI was the most common fully structured interview (11 studies, total n=6272). Among studies that used semistructured, fully structured, and MINI diagnostic interviews, mean sample sizes were 232, 549, and 197, and mean numbers (percentages) with major depression were 32 (14%), 60 (11%), and 37 (19%), respectively (table 2).

**PHQ-9 accuracy by reference standard**

Table 3 and table 4 show comparisons of sensitivity and specificity estimates by reference standard category. A cut-off score of 10 maximized combined sensitivity and specificity among studies using semistructured interviews (sensitivity 0.88, 95% confidence interval 0.83 to 0.92; specificity 0.85, 0.82 to 0.88). Based on cut-off score 10, sensitivity and specificity were 0.70 (0.59 to 0.80) and 0.84 (0.77 to 0.89) for fully structured interviews and 0.77 (0.68 to 0.83) and 0.87 (0.83 to 0.90) for the MINI. Across cut-off scores, specificity estimates were similar across reference standards; however, sensitivity estimates for semistructured interviews were 5-22% higher than for fully structured interviews (median difference 18% at cut-off 10) and 2-15% higher than for the MINI (median difference 11% at cut-off 10). Receiver operating characteristic curves and area under the curve values are shown in supplementary figure B.

Heterogeneity analyses suggested moderate heterogeneity across studies, which improved in some instances when we considered subgroups. Cut-off 10 sensitivity and specificity forest plots are shown in supplementary figure C, with $r^2$ and R values shown in supplementary table C.

Figure 1 shows nomograms of positive and negative predictive values for cut-off score 10 for each reference standard category. For hypothetical values of major depression prevalence of 5-25%, estimates of positive predictive values based on summary sensitivity and specificity values ranged from 24% to 66% for semistructured interviews, 19% to 59% for fully structured interviews, and 24% to 66% for the MINI; estimates of negative predictive values ranged from 96% to 99% for semistructured interviews, 89% to 98% for fully structured interviews, and 92% to 99% for the MINI.

When examined with meta-regression analysis, consistent with our main results, we found that PHQ-9 sensitivity estimates for semistructured interviews were significantly higher than for fully structured interviews or the MINI (supplementary table D). The significant interactions corresponded to differences in sensitivity that across cut-off scores were 4-22% (median 18%) higher for semistructured interviews than for fully structured interviews and 1-16% (median 11%) higher for semistructured interviews than the MINI. Across all cut-off scores, the magnitude of the differences estimated on the basis of meta-regression were within 1% of those estimated using the original two stage bivariate random effects meta-analytic models.

**PHQ-9 accuracy among participants not diagnosed as having or receiving treatment for mental health problem**

Sensitivity and specificity estimates were not statistically significantly different for any reference...
Table 4 | Comparison of sensitivity and specificity estimates among semistructured versus MINI reference standards

<table>
<thead>
<tr>
<th>Cut-off score</th>
<th>Semistructured reference standard*</th>
<th>MINI reference standard†</th>
<th>Difference across reference standards (semistructured minus MINI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity (95% CI) Specificity (95% CI)</td>
<td>Sensitivity (95% CI) Specificity (95% CI)</td>
<td>Sensitivity (95% CI) Specificity (95% CI)</td>
</tr>
<tr>
<td>5</td>
<td>0.98 (0.96 to 0.99) 0.55 (0.49 to 0.60)</td>
<td>0.96 (0.93 to 0.98) 0.57 (0.50 to 0.64)</td>
<td>0.02 (-0.02 to 0.07) -0.02 (-0.14 to 0.11)</td>
</tr>
<tr>
<td>6</td>
<td>0.98 (0.95 to 0.99) 0.63 (0.58 to 0.67)</td>
<td>0.93 (0.87 to 0.97) 0.66 (0.61 to 0.72)</td>
<td>0.05 (-0.01 to 0.12) -0.03 (-0.13 to 0.09)</td>
</tr>
<tr>
<td>7</td>
<td>0.98 (0.94 to 0.99) 0.69 (0.63 to 0.74)</td>
<td>0.90 (0.82 to 0.96) 0.72 (0.66 to 0.78)</td>
<td>0.08 (-0.00 to 0.16) -0.03 (-0.12 to 0.08)</td>
</tr>
<tr>
<td>8</td>
<td>0.95 (0.91 to 0.97) 0.75 (0.71 to 0.79)</td>
<td>0.86 (0.78 to 0.91) 0.78 (0.73 to 0.83)</td>
<td>0.09 (-0.01 to 0.19) -0.03 (-0.11 to 0.06)</td>
</tr>
<tr>
<td>9</td>
<td>0.91 (0.87 to 0.94) 0.80 (0.77 to 0.83)</td>
<td>0.82 (0.72 to 0.88) 0.84 (0.79 to 0.87)</td>
<td>0.09 (-0.02 to 0.22) -0.04 (-0.09 to 0.05)</td>
</tr>
<tr>
<td>10</td>
<td>0.88 (0.83 to 0.92) 0.85 (0.82 to 0.88)</td>
<td>0.77 (0.68 to 0.83) 0.87 (0.83 to 0.90)</td>
<td>0.11 (-0.01 to 0.25) -0.02 (-0.07 to 0.06)</td>
</tr>
<tr>
<td>11</td>
<td>0.84 (0.78 to 0.89) 0.86 (0.86 to 0.91)</td>
<td>0.70 (0.62 to 0.77) 0.90 (0.86 to 0.92)</td>
<td>0.14 (0.01 to 0.30) -0.01 (-0.06 to 0.05)</td>
</tr>
<tr>
<td>12</td>
<td>0.79 (0.73 to 0.83) 0.91 (0.89 to 0.93)</td>
<td>0.65 (0.56 to 0.72) 0.92 (0.89 to 0.94)</td>
<td>0.14 (-0.01 to 0.28) -0.01 (-0.05 to 0.05)</td>
</tr>
<tr>
<td>13</td>
<td>0.70 (0.65 to 0.75) 0.93 (0.91 to 0.95)</td>
<td>0.57 (0.49 to 0.65) 0.94 (0.91 to 0.96)</td>
<td>0.13 (-0.03 to 0.26) -0.01 (-0.04 to 0.04)</td>
</tr>
<tr>
<td>14‡</td>
<td>0.64 (0.58 to 0.70) 0.95 (0.93 to 0.96)</td>
<td>0.49 (0.42 to 0.56) 0.96 (0.93 to 0.97)</td>
<td>0.15 (0.01 to 0.28) -0.01 (-0.04 to 0.03)</td>
</tr>
<tr>
<td>15‡</td>
<td>0.56 (0.50 to 0.62) 0.96 (0.95 to 0.97)</td>
<td>0.42 (0.35 to 0.49) 0.97 (0.95 to 0.98)</td>
<td>0.14 (-0.01 to 0.27) -0.01 (-0.03 to 0.02)</td>
</tr>
</tbody>
</table>

MINI=Mini International Neuropsychiatric Interview.
*Studies n=29; participants n=6725; major depression n=924.
†Studies n=15; participants n=2952; major depression n=549.
‡For these cut-off scores, among studies that used MINI as reference standard, default optimizer in glmer failed, so bobyqa was used instead.

PHQ-9 accuracy among subgroups

For each reference standard category, comparisons of sensitivity and specificity estimates based on bivariate models across PHQ-9 cut-off scores 5-15 among subgroups based on age, sex, country human development index, and participant recruitment setting are shown in supplementary table E, with forest plots shown in supplementary figure D, receiver operating characteristic curves and area under the curve values.

Risk of bias sensitivity analyses

Supplementary table F shows Quality Assessment of Diagnostic Accuracy Studies-2 ratings for each included primary study, and comparisons of PHQ-9 accuracy across individual items for each reference standard category are shown in supplementary table E. For the item on blinding of the reference standard to PHQ-9 results, specificity was significantly greater for studies and participants with high or unclear risk
versus low risk of bias for semistructured interviews but significantly greater for low risk versus high or unclear risk of bias for fully structured interviews and the MINI. For the item on recruiting a consecutive or random sample of participants, specificity was significantly greater for low risk versus high or unclear risk of bias for fully structured interviews and the MINI. We found no other statistically significant differences, and no significant differences were replicated across all reference standards.

**Discussion**

We compared the accuracy of scores on the PHQ-9 for screening to detect major depression, separately, with semistructured diagnostic interviews, fully structured diagnostic interviews (MINI excluded), and the MINI. Based on results from the semistructured interviews, which most closely replicate clinical interviews done by trained professionals, the PHQ-9 was more sensitive than has been reported in previous meta-analyses that combined reference standards. Specificity was similar to previous studies and across reference standards. Based on semistructured interviews, the standard cut-off score of 10 maximized combined sensitivity and specificity. We found evidence from multivariable meta-regression that the PHQ-9 may be more sensitive among older patients than younger patients, but this would not require that a different cut-off score be used. Results did not differ depending on whether studies that did not explicitly exclude patients already diagnosed as having depression were included or excluded. Among studies conducted in primary care settings, approximately half of patients who screened positive on the PHQ-9 had major depression.

**Findings in context**

This is the first meta-analysis that has analyzed diagnostic accuracy for the PHQ-9 separately for different diagnostic interviews. Diagnostic interviews that are used to classify case status for major depression are imperfect reference standards. Semistructured interviews, such as the SCID, most closely approximate an expert diagnosis. They are set up to replicate a guided diagnostic conversation with standardized questions, with the option for interviewers to make additional queries and use clinical judgment to decide whether symptoms are present. Semistructured interviews involve lengthy processes that must be conducted by skilled diagnosticians and, thus, are expensive. Fully structured interviews, such as the CIDI, are designed to replicate as closely as possible expert administered semistructured interviews but are not expected to have the same level of validity and reliability. Fully structured interviews can be administered by lay interviewers and involve fully scripted standardized interview protocols that are read verbatim without additional probes or interpretation. Fully structured interviews are designed to increase reliability with administration by lay interviewers who are not trained to carry out diagnostic interviews independently at the possible cost of validity. The MINI is a specific fully structured interview that was designed to be administered in a fraction of the time compared with other interviews and described by its developers as intentionally overinclusive. Test-retest reliability for diagnosis of current major depression has been reported to be $\kappa = 0.74$ for the SCID (n=51; mean 9 days) and $\kappa = 0.52$ for the CIDI (n=60; mean 2 days).

Consistent with the design features and rigor of each type of diagnostic interview, we previously reported that compared with semistructured interviews, fully structured interviews (excluding the MINI) classify more people with low symptoms as having major depression but fewer people with high symptoms. We also found that the MINI identified approximately twice as many cases as other fully structured interviews. The finding in this study that sensitivity was greater among studies with semistructured than fully structured reference standards is consistent with both the design features and rigor of the different types of diagnostic interviews and with our previous findings. The lower sensitivity among fully structured interviews may have been due to overdiagnosis of major depression among participants with low depressive symptom levels when fully structured interviews were used. In this meta-analysis, most participants (87%) did not have major depression, so misclassification of major depression among participants with subthreshold depressive symptom levels based on fully structured interviews might explain the lower sensitivity compared with semistructured interviews if the PHQ-9 were less...
likely to identify “false positive” classifications based on fully structured interviews. The same logic would apply to the lower sensitivity for the MINI.

Among studies that used semistructured reference standards, sensitivity was also greater than reported in previous traditional meta-analyses, in which studies with semistructured and fully structured reference standards and the MINI were combined without adjustment. Using individual participant data from the 29 studies that used a semistructured interview as the reference standard, we found that at a cut-off score of 10, sensitivity and specificity were 0.88 and 0.85 compared with 0.78 and 0.87 in a 2015 conventional meta-analysis of 34 studies that combined reference standards. In primary care settings, we found sensitivity and specificity of 0.94 and 0.88 (nine studies with a semistructured interview) compared with 0.82 and 0.85 in a 2016 conventional meta-analysis of 20 studies that combined reference standards. For semistructured interviews, prevalence of major depression in our dataset was 14%. Using our cut-off 10 accuracy estimates (sensitivity 0.88, specificity 0.85), the positive predictive value would be only 49%; thus 51% of all positive screens would be false positives. For primary care settings, where accuracy was even higher, prevalence of major depression was 12%. Using our accuracy estimates for cut-off 10 (sensitivity 0.94, specificity 0.88, positive predictive value 52%), 22% of patients in primary care would screen positive at this cut-off score, but only approximately half would be true positives.

To facilitate understanding for clinicians considering use of the PHQ-9 to screen for depression, we have developed a web based tool (depressionscreening100.com/phq). The tool can be used to estimate the expected number of positive screens and true and false screening outcomes based on results from this study.

Clinical implications
Screening for depression in primary care is recommended in the US, but national guidelines from Canada and the UK advise against routine screening for depression. Those guidelines cite the lack of evidence of benefit from well conducted randomized controlled trials, as well as concerns about high false positive rates, overdiagnosis, and substantial resource use and opportunity costs. Well conducted and adequately powered trials designed specifically to assess the effects of depression screening are needed. If screening is to be done clinically on the basis of recommendations in the US, the cut-off score that maximizes sensitivity and specificity is the standard cut-off of 10 or greater. Whether using this standard cut-off score would maximize the likelihood that screening would successfully improve mental health and minimize unnecessary resource use and adverse outcomes if tested in a trial is, however, not known. Ideally, robust trials that are sufficiently powered to evaluate the effects of screening across a range of cut-off scores will be conducted. Clinical trials provide the best possible evidence to inform both the decision on whether depression screening should be implemented as part of routine care and, if so, the thresholds for intervening or what steps might be taken for patients with borderline screening results.

Strengths and limitations of study
This was the first study to use individual participant data meta-analysis to assess diagnostic accuracy of the PHQ-9 or any other depression screening tool. Strengths include the large sample size, the ability to include results from all cut-off scores from all studies (rather than just those published), the ability to examine subgroups of participants, and the ability to assess accuracy separately across reference standards, which had not been done previously. Some limitations should also be considered. Firstly, we were unable to include primary data from 14 of 69 published eligible datasets (20% of eligible datasets and participants), and we restricted our analyses to those with complete data for all variables used in our various analyses (98% of available data). Nevertheless, for all cut-off scores other than 10, our sample was much larger than previous traditional meta-analyses of the PHQ-9. Secondly, despite the large sample size, substantial heterogeneity existed across studies, although it improved in some instances when we considered subgroups. We were not able to do subgroup analyses based on specific medical comorbidities or cultural aspects such as country or language, because comorbidity data were not available for more than half of participants and many countries and languages were represented in few primary studies. However, we were able to compare participant subgroups based on age, sex, country human development index, and participant recruitment setting category, which has not been done previously. Thirdly, although we categorized studies on the basis of the diagnostic interview administered, interviews are sometimes adapted and thus not always used in the way that they were originally designed. Although we coded for qualifications of interviewer for all semistructured interviews as part of our Quality Assessment of Diagnostic Accuracy Studies-2 rating, two studies used interviewers who did not meet typical standards, and approximately half of studies were rated as unclear on this item.

Although our original two stage bivariate random effects meta-analytic models did not find significant differences in accuracy estimates across participant subgroups, our meta-regressions suggested that specificity might be somewhat higher among older participants whether measured continuously or dichotomously. This difference in significance may be due to the differences between the analytical approaches. Whereas statistical significance of the interactions between covariates and accuracy estimates in the meta-regressions were based on parametric standard errors, statistical significance of subgroup comparisons in the two stage bivariate random effects models was based on non-parametric bootstrap methods. Moreover, whereas the meta-regression models provide a within study interpretation, the two
stage bivariate random effects models did not link study clusters across subgroups and thus focused more on between study comparisons.

Conclusions and policy implications
In summary, we found that the sensitivity of PHQ-9 compared with semistructured reference standards was substantially greater than when compared with fully structured reference standards or the MINI. It was also substantially higher than previously reported in conventional meta-analyses that combined reference standards.8 36 The standard cut-off score of 10 or greater maximized combined sensitivity and specificity. However, in primary care, approximately half of patients with positive screens would be false positives if this was used in practice, a concern that has been emphasized by the Canadian Task Force on Preventive Health Care, UK National Screening Committee, and UK National Institute for Health and Care Excellence, given the resources that would be needed for additional assessment and the possibility that some of these patients might be treated without benefit.40 41 43 Future research on the PHQ-9 should ideally be based on semistructured diagnostic interviews, should consider estimating probabilities of depression across the full spectrum of PHQ-9 screening scores (rather than dichotomizing scores at a cut-off), and should combine screening scores with individual characteristics to generate individualized probabilities of major depression.

Contributors: BLevis, AB, BDT, JB, PC, SG, JPM, LAK, DAM, SBP, JIS, and RCZ were responsible for the study conception and design: JB and LAK designed and conducted database searches to identify eligible studies. BDT, DHA, BA, LRH, MB, CHB, PB, GC, MHC, JCN, KC, YC, JMG, JD, JRF, JHF, DF, BG, FGs, CGG, BJH, JH, JPA, HMF, UH, LH, SEH, MHudson, MI, KI, NJ, MPK, KMK, YK, SU, ML, SRL, DBlowe, LM, AM, SMS, TMN, KM, FLO, VP, BWP, PP, AP, KR, AGR, ISS, JS, ASidebottom, ASimning, LS, SCS, PLTT, AT, CMoFOfC, HCW, PAV, JW, MW, KMY, and YZ contributed primary datasets that were included in this study. BLevis, KER, NS, MA, DMR, MJ, TAS, and BDT contributed to data extraction and coding for the meta-analysis. BLevis, AB, BDT, and AWL contributed to the data analysis and interpretation. BLevis, AB, and BDT contributed to drafting the manuscript. All authors provided a critical review and approved the final manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. AB and BDT are the guarantors. The affiliations of the members of the DEPRESSD Collaboration are given in the supplementary materials.


Funding: This study was funded by the Canadian Institutes of Health Research (CIHR, KRS-134297). BLevis was supported by a CIHR Frederick Banting and Charles Best Canada Graduate Scholarship doctoral award. AB and BDT were supported by Fonds de recherche du Québec - Santé (FRQS) researcher salary awards. KER and NS were supported by CIHR Frederick Banting and Charles Best Canada Graduate Scholarship master’s awards. AWL and MA were supported by FRQS Masters Training Awards. BDR was supported by a Vanier Canada Graduate Scholarship. Collection of data for the study by Arolt et al was supported by a project grant from the Health Research Council of New Zealand. Data collection for the study by Ayalon et al was supported from a grant from Lundbeck International. The primary study by Khanshe et al was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary study by Bombardier et al was supported by the Department of Epidemiology, National Institute on Disability, Rehabilitation Research, Spinal Cord Injury Model Systems-University of Washington (grant No H133N060003), Baylor College of Medicine (grant No H133N060003), and University of Michigan (grant No H133N0600012). PB was supported by Australian Research Council Future Fellowships FT130100172. Collection of data for the primary study by Zhang et al was supported by the European Foundation for Study of Diabetes, the Chinese Diabetes Society, Lilly Foundation, Asia Diabetes Foundation, and Liao Wun Yuk Diabetes Memorial Fund. YC received support from NIHM (R24MH701640) and the Centers for Disease Control and Prevention (R48 CE002093). Collection of data for the primary study by Delgadoj et al was supported by grant from St Anne’s Community Services, Leeds, UK. Collection of data for the primary study by Fann et al was supported by grant R01 HD19415 from the US National Institute on Disability and Rehabilitation Research. The primary studies by Amozejogar and by Fiest et al were funded by the Alberta Health Services, the University of Calgary Faculty of Medicine, and the Hotchkiss Brain Institute. The primary study by Fischer et al was funded by the German Federal Ministry of Education and Research (01GI1150). Data for the primary study by Gelaye et al was supported by grant from the NIH (T37 MD001449). Collection of data for the primary study by Gjørgjerd et al was supported by grants from the NIHM (R34 MH72925, KO2 MH66919, P30 DK05456). The primary study by Eal et al was funded by the NIHM (R24 MH65859). Collection of data for the primary study by Hobofl et al was made possible in part by grants from NIHM (R01 MH073687) and the Ohio Board of Regents. BJH received support from a grant awarded by the Research and Development Administration Office, University of Macau (MYRG2015-00109-FSS). Collection of data for the primary study by Harter and KR was supported by the Federal Ministry of Education and Research (grants No 01 GD 9802/4 and 01 GD 0101) and by the Federation of German Pension Insurance Institute. The primary study by Hides et al was funded by the Perpetual Trustees, Flora and Frank Leith Charitable Trust, Jack Brockhoff Foundation, Groove & Settlement, Sunshine Foundation, and Danks Trust. The primary study by Henkel et al was funded by the German Ministry of Research and Education. Data for the study by Razkov et al was collected by the Canadian Scleroderma Research Group, which was funded by the CIHR (FRN 83158). The Scleroderma Society of Canada, the Scleroderma Society of Ontario, the Scleroderma Society of Saskatchewan, Sclérodémie Québec, the Cure Scleroderma Foundation, Inova Diagnostics Inc, Euroimmun, FRQS, the Canadian Arthritis Network, and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montreal, QC. M Hudson was supported by a FRQS Senior Investigator Award. Collection of data for the primary study by Hyphantis et al was supported by grant from the National Strategic Reference Framework, European Union, and the Greek General Secretariat for Research and Technology, National Strategic Reference Framework, R&D (ARISTDA-ABREVIA, 1259). The primary study by Inagaki et al was supported by the Ministry of Health, Labour and Welfare, Japan. N was supported by a Canada Research Chair in Neurological Health Services Research. Collection of data for the primary study by Iekeda et al was supported by National Health and Medical Research Council (grant No 1002160) and Safe Work Australia. KMK was supported by funding from a Australian National Health and Medical Research Council fellowship (grant No 1088313). The primary study by Lamers et al was funded by the Netherlands Organisation for Health Research and Development (grant No 955-03-047). The primary study by Liu et al was funded by a grant (M-288) from Tehran University of Medical Sciences. Collection of data for the primary study by Lottakul et al was supported by the Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (grant No 49086). BLowe received research grants from Pfizer, Germany, and from the medical faculty of the University of Heidelberg, Germany (project 121/2000) for the study by Grafe et
al. The primary study by Mohd Sidik et al was funded under the Research University Grant Scheme from Universiti Putra Malaysia, Malaysia, and the Postgraduate Research Student Support Accounts of the University of Auckland, New Zealand. The primary study by Santos et al was funded by the National Program for Centers of Excellence (PRONEX/FAPERGS/CNPq, Brazil). The primary study by Muramatsu et al was supported by an educational grant from Pfizer US Pharmaceutical Inc. Collection of data for the study by BWP was provided by NIMH (R34MH084673). The primary studies by Osano et al were funded by Reitona de Pesquisa da Universidade de Sao Paulo (MCT 09.1.01689.17.17) and Banco Santander (grant No 10.01232.17.9). FLO was supported by Productivity Grants (PQ-CNPG-2 number 301321/2016-7). The primary study by Picardi et al was supported by funds for current research from the Italian Ministry of Health. PP was supported by a grant from the Belgian Ministry of Public Health and Social Affairs and a restricted grant from Pfizer Belgium. JS was supported by funding from Universiti Sains Malaysia. The primary study by Rooney et al was funded by the UK National Health Service Lothian Neuro-Oncology Endowment Fund. The primary study by Sidebottom et al was funded by a grant from the United States Department of Health and Human Services, Health Resources and Services Administration (grant No R44MC07840). Sinning et al’s research was supported in part by grants from the NIH (T32 GM07356), Agency for Healthcare Research and Quality (R36 HS018246), NIMH (R24 MH071504), and the National Center for Research Resources (TL1 RR024135). LS received PhD scholarship funding from the University of Melbourne. Collection of data for the studies by Turner et al were funded by a bequest from Jennie Thomas through the Hunter Medical Research Institute. The study by van Steenbergen-Weijenberg et al was funded by Innovatiefonds Zorgverzekeraars. PV was supported by the Fund for Innovation and Competitiveness of the Chilean Ministry of Economy, Development and Tourism, through the Millennium Scientific Initiative (grant No IS130005). Collection of data for the primary study by Williams et al was supported by an NIMH grant to UM (R01-MH069666). The primary study by Thombs et al was done with data from the Heart and Soul Study (PI Mary Whooley). The Heart and Soul Study was funded by the Department of Veterans Epidemiology Ment Research Program, the Department of Veterans Affairs Health Services Research and Development service, the National Heart Lung and Blood Institute (R01 HL079235), the American Federation for Ageing Research, the Robert Wood Johnson Foundation, and the Ischemia Research and Education Foundation. The primary study by Twist et al was funded by the UK National Institute for Health Research under its Programme Grants for Applied Research Programme (grant reference No RP-PG-0606-1142). The study by Wittkamp et al was funded by the Netherlands Organization for Health Research and Development (ZonMw) Mental Health Program (No 100.003.005 and 100.002.021) and the Academic Medical Center/University of Amsterdam. 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 of the 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/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work other than that described above; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years with the following exceptions: NJ and SP received a grant, outside the submitted work, from the University of Calgary Hotchkiss Brain Institute, which was jointly funded by the Institute and Pfizer; Pfizer was the original sponsor of the development of the PHQ-9, which is now in the public domain. JNC is a steering committee member or consultant of Astra Zeneca, Bayer, Lilly, MSD, and Pfizer and has received sponsorships and honorariums for giving lectures and providing consultancy, and her affiliated institution has received research grants from these companies. LH was an advisory board member for Lundbeck and Servier, a consultant for Bayer Pharma, and a speaker for Roche Pharma and Servier and has received personal fees from Janssen, all outside the submitted work; MI has received a grant from Novartis Pharma and personal fees from Meij, Mohchida, Takeda, Novartis, Yoshitomi, Pfizer, Eisai, Onkoku, MSD, Technronics, and Sumitomo Dainippon, all outside of the submitted work; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: As this study involved secondary analysis of anonymized previously collected data, the Research Ethics Committee of the Jewish General Hospital declared that this project did not need research ethics approval. However, for each included dataset, the authors confirmed that the original study received ethics approval and that all patients provided informed consent.

Transparency declaration: 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.

Data sharing: Requests to access data should be made to the corresponding author at brett.thombs@mccill.ca.

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license its derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

1 Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care? BMJ 2014;348:g1253. doi:10.1136/bmj.g1253
10 Thombs BD, Hsu E, El-Baalbaki GA, et al. 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
15 Brugha TS, Jenkins R, Taib 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/S003329170100418A
16 Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic


Supplementary materials