IMPORTANCE The Patient Health Questionnaire depression module (PHQ-9) is a 9-item self-administered instrument used for detecting depression and assessing severity of depression. The Patient Health Questionnaire-2 (PHQ-2) consists of the first 2 items of the PHQ-9 (which assess the frequency of depressed mood and anhedonia) and can be used as a first step to identify patients for evaluation with the full PHQ-9.

OBJECTIVE To estimate PHQ-2 accuracy alone and combined with the PHQ-9 for detecting major depression.

DATA SOURCES MEDLINE, MEDLINE In-Process & Other Non-Indexed Citations, PsycINFO, and Web of Science (January 2000-May 2018).

STUDY SELECTION Eligible data sets compared PHQ-2 scores with major depression diagnoses from a validated diagnostic interview.

DATA EXTRACTION AND SYNTHESIS Individual participant data were synthesized with bivariate random-effects meta-analysis to estimate pooled sensitivity and specificity of the PHQ-2 alone among studies using semistructured, fully structured, or Mini International Neuropsychiatric Interview (MINI) diagnostic interviews separately and in combination with the PHQ-9 vs the PHQ-9 alone for studies that used semistructured interviews. The PHQ-2 score ranges from 0 to 6, and the PHQ-9 score ranges from 0 to 27.

RESULTS Individual participant data were obtained from 100 of 136 eligible studies (44 318 participants; 4572 with major depression [10%]; mean [SD] age, 49 [17] years; 59% female). Among studies that used semistructured interviews, PHQ-2 sensitivity and specificity (95% CI) were 0.91 (0.88-0.94) and 0.67 (0.64-0.71) for cutoff scores of 2 or greater and 0.72 (0.67-0.77) and 0.85 (0.83-0.87) for cutoff scores of 3 or greater. Sensitivity was significantly greater for semistructured vs fully structured interviews. Specificity was not significantly different across the types of interviews. The area under the receiver operating characteristic curve was 0.88 (0.86-0.89) for semistructured interviews, 0.82 (0.81-0.84) for fully structured interviews, and 0.87 (0.85-0.88) for the MINI. There were no significant subgroup differences. For semistructured interviews, sensitivity for PHQ-2 scores of 2 or greater followed by PHQ-9 scores of 10 or greater was not significantly different than PHQ-9 scores of 10 or greater alone (0.82 [0.76-0.86]). Specificity for the combination was significantly but minimally higher (0.87 [0.84-0.89] vs 0.85 [0.82-0.87]). The area under the curve was 0.90 (0.89-0.91). The combination was estimated to reduce the number of participants needing to complete the full PHQ-9 by 57% (56%-58%).

CONCLUSIONS AND RELEVANCE In an individual participant data meta-analysis of studies that compared PHQ scores with major depression diagnoses, the combination of PHQ-2 (with cutoff ≥2) followed by PHQ-9 (with cutoff ≥10) had similar sensitivity but higher specificity compared with PHQ-9 cutoff scores of 10 or greater alone. Further research is needed to understand the clinical and research value of this combined approach to screening.
n depression screening, questionnaires are used to identify patients with scores above a cutoff threshold for evaluation to determine whether depression is present. One strategy is to administer a brief screening tool followed by a longer tool for positive screens. The Patient Health Questionnaire–2 (PHQ-2), which consists of the first 2 items (depressed mood and anhedonia) of the Patient Health Questionnaire–9 (PHQ-9), has been recommended as a prescreen prior to administering remaining PHQ-9 items (Table 1). A 2016 aggregate-data meta-analysis on PHQ-2 accuracy included 21 published studies of the PHQ-2; however, it did not include PHQ-2 data from an additional 37 studies of the PHQ-9, except for clinical setting, subgroup results were not reported in primary studies and not evaluated; all primary studies were synthesized regardless of the diagnostic interview used, despite differences in their likelihood of classifying major depression, and PHQ-2 accuracy was not evaluated in combination with the PHQ-9, as typically used in practice. Two primary studies have evaluated the PHQ-2 and PHQ-9 combination and produced inconsistent results; one examined score cutoffs for PHQ-2 of 2 or greater and for PHQ-9 of 10 or greater in older community-dwelling adults, and the other examined score cutoffs for PHQ-2 of 2 or greater and for PHQ-9 of 6 or greater in patients with acute coronary syndrome.

The objectives of this meta-analysis of individual participant data were to evaluate PHQ-2 screening accuracy in adults (1) among studies that used different types of reference standards separately; (2) among participants verified as not diagnosed or in treatment vs all participants and by subgroups based on age, sex, country Human Development Index, and recruitment setting; and (3) alone and in combination with the PHQ-9 vs the PHQ-9 alone.

### Methods

We published a protocol and registered in PROSPERO (CRD42014010673). Results were reported per PRISMA-DTA and PRISMA-IPD. Previous publications reported PHQ-8 and PHQ-9 accuracy. Individual prediction models described in the protocol will be developed in future studies. Analysis of the PHQ-2 and PHQ-9 combination was not prespecified. This study involved analysis of previously collected deidentified data, and included studies were required to have obtained ethics approval and informed consent; thus, the research ethics committee of the Jewish General Hospital determined that ethics approval was not required.

### Study Eligibility

Studies were sought with data sets that (1) included PHQ-2 scores or item data to calculate PHQ-2 scores; (2) included current major depressive disorder or major depressive episode classification based on Diagnostic and Statistical Manual of Mental Disorders (DSM) or International Classification of Diseases (ICD) criteria and a validated diagnostic interview; (3) administered the PHQ and diagnostic interview within a 2-week period because diagnostic criteria include only symptoms from the last 2 weeks; (4) included participants 18 years and older not recruited from school or university settings; and (5) did not recruit participants only from psychiatric settings or with depression symptoms because screening is done to identify people not suspected of having depression. In data sets where only some participants were eligible, we included only those participants. There were no language restrictions.

### Database Searches and Study Selection

The database search was designed by a medical librarian and peer-reviewed and included MEDLINE, MEDLINE In-Process & Other Non-Indexed Citations via Ovid, PsycINFO, and Web of Science (January 1, 2000-May 9, 2018) (eMethods 1 in the Supplement). We searched from 2000 because the PHQ-9 was published in 2001. We reviewed review articles and queried contributing authors about nonpublished studies or studies not identified by the search. We uploaded results into RefWorks (RefWorks-COS; Bethesda, Maryland), removed duplicates, then uploaded references into DistillerSR (Evidence Partners; Ottawa, Ontario, Canada).

Titles and abstracts were independently reviewed by varying pairs of 2 investigators. If 1 identified a study as potentially eligible, the full text was reviewed by pairs of 2 investigators independently. Any differences were resolved by consensus, with a third investigator consulted if necessary.

We conducted a literature search on April 6, 2020, to seek eligible published results that could be included. No studies published since the original search provided results for PHQ-2 and PHQ-9 combined.

### Data Contribution, Extraction, and Synthesis

We emailed corresponding authors of studies with eligible data sets at least 3 times, as necessary, to invite them to contribute data sets. If there was no response, we emailed coauthors and attempted contact by telephone.

Country, recruitment setting (nonmedical, primary care, inpatient, outpatient specialty), and diagnostic interview were extracted from published reports by 2 investigators independently, with disagreements resolved by consensus. Countries were categorized as having very high, high, or low-medium development based on the United Nation’s 2019 Human Development Index. Individual participant records included sex, age, major depression status, current mental health diagnosis or treatment, and PHQ-2 and PHQ-9 total and item scores.
PHQ-9 items reflect the 9 DSM symptoms of major depression; PHQ-2 items reflect depressed mood and anhedonia. We prioritized major depressive episode over major depressive disorder, if both were provided, because screening attempts to detect episodes, and we prioritized DSM over ICD. For 4 studies with multiple recruitment settings, setting was coded by participants. When primary studies provided sampling weights, we used those weights. If weighting should have been done but was not, we used inverse selection probability weights. If all study participants with scores above a threshold but only a random subset of 50% below the threshold received a diagnostic interview, for instance, those above the threshold received a weight of 1 and those below received a weight of 2.

For each included data set, we attempted to replicate published participant characteristics and accuracy results. We worked with primary study investigators to resolve any discrepancies.

**Risk of Bias Assessment**
Risk of bias was assessed with the Quality Assessment of Diagnostic Accuracy Studies–2 tool (QUADAS-2; eMethods 2 in the Supplement).28 This was done by 2 investigators independently with discrepancies resolved by consensus, involving a third investigator, if necessary.

**Statistical Analyses**
The PHQ-2 score ranges from 0 to 6, and the PHQ-9 score ranges from 0 to 27. We estimated sensitivity and specificity for all possible PHQ-2 cutoffs (scores 1-6) by reference standard type separately: semistructured diagnostic interviews; fully structured diagnostic interviews, excluding the Mini International Neuro-psychiatric Interview (MINI)29,30; and the MINI. We did this because, controlling for depressive symptom scores, the Composite International Diagnostic Interview (CIDI),31 the most commonly used fully structured interview, may classify more participants with low-level symptoms as depressed, but fewer participants with higher-level symptoms, than semistructured interviews.11-13 The MINI may classify more participants as depressed.31-33 This is consistent with interview designs. Semistructured interviews are intended for administration by experienced diagnosticians, require clinical judgment, and allow question rephrasing and probes. Fully structured interviews are designed for lay interviewer administration and are fully scripted with no deviation allowed. They are intended to achieve standardization but may sacrifice accuracy.32-35 The MINI was designed for rapid administration and to be overinclusive.29,30

Within each reference standard category, we conducted subgroup analyses. We estimated sensitivity and specificity among participants who could be verified as not currently diagnosed or receiving mental health treatment vs all participants. This is because some primary studies included people already diagnosed or receiving treatment, but those participants would not be screened in practice. We estimated sensitivity and specificity by age (<60, ≥60 years), sex, country, Human Development Index, and recruitment setting.

Among studies that used a semistructured interview, we evaluated accuracy of the PHQ-2 and PHQ-9 combination based on commonly used cutoffs.8,20 We compared sensitivity and specificity for PHQ-2 scores of 2 or greater and 3 or greater alone and combined with PHQ-9 scores of 10 or greater vs PHQ-9 scores of 10 or greater alone. In each scenario, we calculated the number of participants who scored above the PHQ-2 threshold and, in practice, would need to complete the full PHQ-9. For these analyses, we excluded studies and participants without PHQ-9 scores. In additional analyses, we compared sensitivity and specificity for PHQ-2 scores of 2 or greater in combination with PHQ-9 cutoff scores of 5 to 15 vs PHQ-9 alone at cutoff scores of 5 to 15.

In all meta-analyses, for all cutoff scores separately, we fit bivariate random-effects models using Gauss–Hermite quadrature.36 This 2-stage approach simultaneously models sensitivity and specificity, accounting for the correlation between them and within-study precision estimates. Within each reference standard category, we constructed empirical receiver operating characteristic plots and calculated area under the curve (AUC). To compare results between subgroups and for the PHQ-2 and PHQ-2 and PHQ-9 combination vs PHQ-9 alone, we estimated sensitivity and specificity differences and constructed confidence intervals for differences via the cluster bootstrap.
Sensitivity and Specificity

Sensitivity and specificity estimates were not significantly different for participants verified as not currently diagnosed or receiving mental health treatment compared with all participants across reference standard categories. Among other subgroup comparisons, there were no statistically significant or substantive differences that replicated across cutoffs and reference standard categories (eTable 4; forest plots: eFigure 2; τ² and R values: eTable 3 in the Supplement).

Comparison of PHQ-2, PHQ-2 in Combination With PHQ-9 for Screening to Detect Major Depression

Based on 44 studies that used a semistructured reference standard and provided both PHQ-2 and PHQ-9 scores, compared with PHQ-9 scores of 10 or greater alone, all strategies resulted

Results

Search Results and Data Set Inclusion

The database search identified 9674 unique citations, of which 9198 were excluded after title and abstract review and 289 after full-text review, leaving 187 eligible articles with 131 unique data sets. Of these, 100 (76%) contributed data sets with PHQ-9 scores, PHQ-2 scores, or both. Authors of included studies contributed data from 5 additional unpublished studies, for a total of 105 data sets. Five data sets with PHQ-9 total scores did not have item data necessary to calculate PHQ-2 scores and were excluded. Thus, 100 data sets (44318 participants; 4572 cases [10%]; mean [SD] age, 49 [17] years; 59% female) were included (Figure 1). eTable 1 in the Supplement shows study characteristics of included studies and eligible studies that did not provide data. Not counting the 5 unpublished studies, of 54 633 participants in 131 eligible published studies, we included 43 787 participants (80%) from 95 published studies (73%).

PHQ-2 Sensitivity and Specificity

Among studies with a semistructured interview, sensitivity and specificity for PHQ-2 scores of 2 or greater were 0.91 (95% CI, 0.88-0.94) and 0.67 (95% CI, 0.64-0.71); for PHQ-2 scores of 3 or greater, sensitivity and specificity were 0.72 (95% CI, 0.67-0.77) and 0.85 (95% CI, 0.83-0.87), respectively. Across cutoffs, sensitivity with semistructured interviews was 0.04 (95% CI, 0.01-0.08) to 0.20 (95% CI, 0.10-0.28) higher than with fully structured interviews (significantly higher for cutoffs 1-6) and 0.02 (95% CI, 0.00-0.04) to 0.05 (95% CI, -0.04-0.13) higher than with the MINI (not significantly different at any cutoff); specificity was not significantly different across reference standard types (Table 3; eFigure 1 in the Supplement). The AUC was 0.88 (95% CI, 0.86-0.89) for semistructured interviews, 0.82 (95% CI, 0.81-0.84) for fully structured diagnostic interviews, and 0.87 (95% CI, 0.85-0.88) for the MINI.

There was moderate heterogeneity. For cutoffs 2 to 3, the τ² values ranged from 0.47 to 1.29 for sensitivity and 0.27 to 0.78 for specificity, while R values ranged from 2.22 to 3.50 for sensitivity and 3.47 to 9.30 for specificity. Forest plots are shown in eFigure 2 and τ² and R values in eTable 3 in the Supplement.

Subgroup Analyses

Sensitivity and specificity estimates were not significantly different for participants verified as not currently diagnosed or receiving mental health treatment compared with all participants across reference standard categories. Among other subgroup comparisons, there were no statistically significant or substantive differences that replicated across cutoffs and reference standard categories (eTable 4; forest plots: eFigure 2; τ² and R values: eTable 3 in the Supplement).
in substantially reduced sensitivity or specificity, except PHQ-2 scores of 2 or greater in combination with PHQ-9 scores of 10 or greater. For this combination, sensitivity was 0.82 (95% CI, 0.76-0.86) vs 0.86 (95% CI, 0.80-0.90) (not statistically significant) and specificity was slightly higher (0.87 [95% CI, 0.84-0.89] vs 0.85 [95% CI, 0.82-0.87]) (statistically significant; Table 4; eTable 5 in the Supplement; Figure 2). The AUC was 0.90 (95% CI, 0.89-0.91). Nomograms of positive and negative predictive values are shown in eFigure 3 in the Supplement. Using PHQ-2 scores of 2 or greater in combination with other PHQ-9 cutoffs (5-9, 11-15) resulted in lower combined sensitivity and specificity compared with PHQ-2 scores of 2 or greater with PHQ-9 scores of 10 or greater (eTable 6 in the Supplement).

With PHQ-2 scores of 2 or greater then PHQ-9 scores of 10 or greater, 43% (95% CI, 42%-44%) of participants had positive PHQ-2 screens and would have needed to complete the full PHQ-9 in practice; 23% (95% CI, 22%-24%) of all participants would have had a positive PHQ-9 screen and needed further mental health assessment compared with 25% (95% CI, 24%-26%) for PHQ-9 scores of 10 or greater alone and 43% (95% CI, 42%-44%) for PHQ-2 scores of 2 or greater alone.
Table 2. Participant Data by Subgroup

<table>
<thead>
<tr>
<th>Participant subgroup</th>
<th>Semistructured diagnostic interviews, No. (%)</th>
<th>Fully structured diagnostic interviews, No. (%)</th>
<th>MINI, No. (%)</th>
<th>All interviews, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Studies</td>
<td>Participants in category</td>
<td>Major depression</td>
<td>Studies</td>
</tr>
<tr>
<td>All participants</td>
<td>48</td>
<td>11 703 (100)</td>
<td>1 538 (13)</td>
<td>20</td>
</tr>
<tr>
<td>Subset of participants verified to not currently be diagnosed or receiving treatment for a mental health problem&lt;sup&gt;b&lt;/sup&gt;</td>
<td>25</td>
<td>3 708 (32)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>527 (14)</td>
<td>5</td>
</tr>
<tr>
<td>Age, y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;60</td>
<td>46</td>
<td>7 767 (67)</td>
<td>1 118 (14)</td>
<td>20</td>
</tr>
<tr>
<td>≥60</td>
<td>43</td>
<td>3 888 (33)</td>
<td>415 (11)</td>
<td>16</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>48</td>
<td>7 287 (62)</td>
<td>1 054 (14)</td>
<td>20</td>
</tr>
<tr>
<td>Men</td>
<td>41</td>
<td>4 408 (38)</td>
<td>484 (11)</td>
<td>18</td>
</tr>
<tr>
<td>Country Human Development Index&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very high</td>
<td>37</td>
<td>9 156 (78)</td>
<td>994 (11)</td>
<td>16</td>
</tr>
<tr>
<td>High</td>
<td>8</td>
<td>1 957 (17)</td>
<td>356 (18)</td>
<td>9</td>
</tr>
<tr>
<td>Low-medium</td>
<td>3</td>
<td>590 (5)</td>
<td>188 (32)</td>
<td>4</td>
</tr>
<tr>
<td>Recruitment setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonmedical care&lt;sup&gt;e&lt;/sup&gt;</td>
<td>2</td>
<td>567 (5)</td>
<td>105 (19)</td>
<td>4</td>
</tr>
<tr>
<td>Primary care</td>
<td>15</td>
<td>4 569 (39)</td>
<td>667 (15)</td>
<td>7</td>
</tr>
<tr>
<td>Inpatient specialty care</td>
<td>10</td>
<td>2 019 (17)</td>
<td>184 (9)</td>
<td>2</td>
</tr>
<tr>
<td>Outpatient specialty care</td>
<td>23</td>
<td>4 548 (39)</td>
<td>582 (13)</td>
<td>7</td>
</tr>
</tbody>
</table>

Abbreviation: MINI, Mini International Neuropsychiatric Interview.

<sup>a</sup>The MINI is a very brief fully structured diagnostic interview that was designed for rapid administration by lay interviewers and intended to be overinclusive.

<sup>b</sup>This row contains the subset of participants who could be verified to not currently be diagnosed or receiving treatment for a mental health problem without a diagnosis or treatment were included, whereas those already diagnosed or receiving treatment were excluded.

<sup>c</sup>Percentage refers to percentage of all participants within semistructured, fully structured, MINI, or all interviews who could be verified to not currently be diagnosed or receiving treatment.

<sup>d</sup>Based on Human Development Report 2019. The Human Development Index is a composite index comprised of indicators of life expectancy, education, and per-capita income. In 2019, very-high human development countries include the top 59 countries; high included countries rated 60 to 112; medium, 113 to 151; and low, 152 to 189 (http://hdr.undp.org/en/composite/HDI).

<sup>e</sup>Nonmedical care recruitment included general community samples, as well as samples of older adults (2 studies), domestic workers (1 study), individuals in countries exposed to war (1 study), drug users (1 study), and employees on sickness leave (1 study).
Table 3. Comparison of PHQ-2 Sensitivity and Specificity Estimates Among Semistructured, Fully Structured, and MINI Reference Standards

<table>
<thead>
<tr>
<th>Cutoff score</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98 (0.96 to 0.99)</td>
<td>0.46 (0.42 to 0.51)</td>
<td>0.93 (0.88 to 0.96)</td>
<td>0.48 (0.38 to 0.58)</td>
<td>0.96 (0.94 to 0.98)</td>
<td>0.48 (0.43 to 0.53)</td>
<td>0.04 (0.01 to 0.08)</td>
<td>-0.02 (-0.10 to 0.08)</td>
<td>0.02 (0.00 to 0.04)</td>
<td>-0.01 (-0.07 to 0.04)</td>
</tr>
<tr>
<td>2</td>
<td>0.91 (0.88 to 0.94)</td>
<td>0.67 (0.64 to 0.71)</td>
<td>0.82 (0.75 to 0.87)</td>
<td>0.71 (0.63 to 0.77)</td>
<td>0.89 (0.84 to 0.93)</td>
<td>0.68 (0.64 to 0.73)</td>
<td>0.10 (0.03 to 0.18)</td>
<td>-0.03 (-0.09 to 0.04)</td>
<td>0.02 (-0.02 to 0.09)</td>
<td>-0.01 (-0.06 to 0.04)</td>
</tr>
<tr>
<td>3</td>
<td>0.72 (0.67 to 0.77)</td>
<td>0.85 (0.83 to 0.87)</td>
<td>0.53 (0.44 to 0.62)</td>
<td>0.89 (0.84 to 0.92)</td>
<td>0.69 (0.62 to 0.75)</td>
<td>0.87 (0.84 to 0.90)</td>
<td>0.19 (0.08 to 0.29)</td>
<td>-0.04 (-0.07 to 0.00)</td>
<td>0.03 (-0.06 to 0.11)</td>
<td>-0.02 (-0.05 to 0.02)</td>
</tr>
<tr>
<td>4</td>
<td>0.55 (0.50 to 0.61)</td>
<td>0.93 (0.91 to 0.94)</td>
<td>0.36 (0.30 to 0.43)</td>
<td>0.94 (0.92 to 0.96)</td>
<td>0.50 (0.44 to 0.56)</td>
<td>0.94 (0.91 to 0.96)</td>
<td>0.20 (0.10 to 0.28)</td>
<td>-0.01 (-0.03 to 0.01)</td>
<td>0.05 (-0.04 to 0.13)</td>
<td>-0.01 (-0.03 to 0.01)</td>
</tr>
<tr>
<td>5</td>
<td>0.35 (0.31 to 0.40)</td>
<td>0.97 (0.96 to 0.98)</td>
<td>0.21 (0.16 to 0.26)</td>
<td>0.98 (0.97 to 0.99)</td>
<td>0.30 (0.25 to 0.36)</td>
<td>0.98 (0.97 to 0.98)</td>
<td>0.14 (0.06 to 0.21)</td>
<td>-0.01 (-0.02 to 0.01)</td>
<td>0.05 (-0.03 to 0.13)</td>
<td>-0.01 (-0.01 to 0.01)</td>
</tr>
<tr>
<td>6</td>
<td>0.23 (0.19 to 0.27)</td>
<td>0.99 (0.98 to 0.99)</td>
<td>0.13 (0.09 to 0.17)</td>
<td>0.99 (0.98 to 0.99)</td>
<td>0.18 (0.15 to 0.22)</td>
<td>0.99 (0.99 to 0.99)</td>
<td>0.10 (0.04 to 0.16)</td>
<td>0.00 (-0.01 to 0.00)</td>
<td>0.05 (-0.02 to 0.01)</td>
<td>0.00 (-0.01 to 0.00)</td>
</tr>
</tbody>
</table>

Abbreviations: AUC, area under the curve; MINI, Mini International Neuropsychiatric Interview; PHQ-2, Patient Health Questionnaire–2.

Table 4. Comparison of Sensitivity and Specificity Estimates and Number of Participants Requiring Full PHQ-9 for PHQ-2 Alone, PHQ-2 in Combination With PHQ-9, and PHQ-9 Alone Among 44 Studies (No. of Participants = 10 627; No. of Participants With Major Depression = 1361) That Used a Semistructured Reference Standard and Had Both PHQ-2 and PHQ-9 Item Scores Available

<table>
<thead>
<tr>
<th>Screening strategy</th>
<th>PHQ-2 score ≥2 alone</th>
<th>PHQ-2 score ≥3 alone</th>
<th>PHQ-2 score ≥2 then PHQ-9 score ≥10</th>
<th>PHQ-2 score ≥3 then PHQ-9 score ≥10</th>
<th>PHQ-9 score ≥10 alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ-2 Administered</td>
<td>10 627</td>
<td>10 627</td>
<td>10 627</td>
<td>10 627</td>
<td>10 627</td>
</tr>
<tr>
<td>Positive screens</td>
<td>4529 (42.6)</td>
<td>2650 (24.9)</td>
<td>4529 (42.6)</td>
<td>2650 (24.9)</td>
<td>10 627 (100.0)</td>
</tr>
<tr>
<td>PHQ-9 Administered</td>
<td>4529 (42.6)</td>
<td>2650 (24.9)</td>
<td>10 627 (100.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive screens</td>
<td>2461 (23.2)</td>
<td>1946 (18.3)</td>
<td>2655 (25.0)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity and specificity (95% CI)

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>0.92 (0.88 to 0.95)</th>
<th>0.72 (0.67 to 0.77)</th>
<th>0.82 (0.76 to 0.86)</th>
<th>0.70 (0.64 to 0.75)</th>
<th>0.86 (0.80 to 0.90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity</td>
<td>0.67 (0.63 to 0.70)</td>
<td>0.85 (0.83 to 0.87)</td>
<td>0.87 (0.84 to 0.89)</td>
<td>0.91 (0.89 to 0.93)</td>
<td>0.85 (0.82 to 0.87)</td>
</tr>
</tbody>
</table>

Difference in accuracy estimates (each strategy - PHQ-9 alone) (95% CI)

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>0.06 (0.01 to 0.11)</th>
<th>-0.13 (-0.20 to -0.09)</th>
<th>-0.04 (-0.09 to 0.01)</th>
<th>-0.16 (-0.23 to -0.12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity</td>
<td>-0.18 (-0.21 to -0.16)</td>
<td>0.01 (-0.02 to 0.03)</td>
<td>0.02 (0.00 to 0.03)</td>
<td>0.06 (0.04 to 0.08)</td>
</tr>
</tbody>
</table>

Abbreviation: PHQ, Patient Health Questionnaire.

Risk of Bias Sensitivity Analyses

eTable 7 in the Supplement shows QUADAS-2 ratings for individual signaling items and risk of bias domains for included primary studies. Among 400 total domain ratings (4 per included study), 131 (33%) were coded as having low risk of bias, 253 (63%) as having an unclear risk, 11 (3%) as having a high risk, and 5 (1%) as varying across participants within a study. Three of 48 studies (6%) that used a semistructured interview,
6 of 20 studies (30%) with a fully structured interview, and 9 of 32 studies (28%) with a MINI reference standard had low risk of bias across all 4 domains.

PHQ-2 accuracy comparisons across QUADAS-2 items within reference standard categories are shown in eTable 4 in the Supplement. No statistically significant differences were found that replicated across cutoffs for any reference standard category.

Discussion

In this individual participant data meta-analysis of 44 studies that used semistructured diagnostic interviews to classify depression, sensitivity using the combination of PHQ-2 (cutoff ≥2) and PHQ-9 (cutoff ≥10) was not significantly different than using the full PHQ-9 (cutoff ≥10) for all participants. Specificity for the combination was significantly, though minimally, higher. The combination approach was estimated to reduce the number of participants needing to do the full PHQ-9 by 57% (95% CI, 56%-58%). Compared with the PHQ-9 alone, the PHQ-2 alone resulted in statistically significant lower sensitivity or specificity, depending on the cutoff score.

Consistent with previous findings with the PHQ-9,20 PHQ-2 sensitivity was highest compared with semistructured interviews, which most closely replicate clinical interviews by trained professionals, and lower compared with fully structured interviews and the MINI, although differences compared with the MINI were small and not statistically significant. Specificity estimates were not significantly different across reference standards. There were no significant accuracy differences between subgroups that replicated across reference standard categories, although some subgroups had limited numbers of participants and cases.

The finding that PHQ-2 sensitivity was greater when compared with semistructured rather than fully structured interviews may have occurred because fully structured interviews are designed for reliability at the cost of validity.12-35 Previous studies found that among participants with low-level depressive symptoms, fully structured interviews may classify more participants as having major depression than semistructured interviews but fewer among participants with high-level symptoms.11,12 In the present meta-analysis, most participants did not have major depression. Thus, misclassification of major depression among participants with subthreshold depressive symptoms based on fully structured interviews might explain the lower sensitivity compared with semistructured interviews.

Among studies with semistructured interviews, PHQ-2 sensitivity and specificity were generally similar to estimates reported in a previous aggregate-data meta-analysis that combined reference standards without adjustment.8 Using individual participant data from 48 studies with semistructured interviews in the present study, sensitivity and specificity were, respectively, 0.91 and 0.67 for cutoff scores of 2 or greater and 0.72 and 0.85 for cutoff scores of 3 or greater compared with 0.91 and 0.70 for cutoff scores of 2 or greater (17 studies) and 0.76 and 0.87 for cutoff scores of 3 or greater (19 studies) in the previous meta-analysis. This differed from a meta-analysis of PHQ-9 individual participant data,20 in which, among studies that used a semistructured interview, sensitivity at the standard cutoff score of 10 or greater was substantially greater than reported in a previous aggregate-data meta-analysis that combined reference standards.8,20

No previous meta-analysis and only 2 primary studies14,15 have evaluated the PHQ-2 in combination with the PHQ-9. The 2 primary studies, however, reported results using different cutoff combinations and generated estimates of sensitivity and specificity that differed among older community-dwelling adults (N = 378; sensitivity = 0.81, specificity = 0.89) and patients with coronary artery disease (N = 1024, sensitivity = 0.75, specificity = 0.84). Using individual participant data from 44 primary studies with semistructured interviews in the present study and standard cutoffs, which maximized combined sensitivity and specificity, sensitivity (0.82) for PHQ-2 scores of 2 or greater followed by PHQ-9 scores of 10 or greater was not significantly different from PHQ-9 scores of 10 or greater alone, and specificity (0.87) was significantly better, though minimally. Assuming that screening procedures allow for quick calculation of PHQ-2 scores before presenting remaining PHQ-9 items (eg, electronic administration), the combination could improve efficiency.

Routine screening for depression in primary care has been recommended in the United States.8 National guidelines from Canada and the United Kingdom, however, recommended...
against screening due to the lack of direct trial evidence of benefit and concerns about harms and consumption of health care resources.\textsuperscript{49,52} Well-conducted trials that compare screening vs no screening are needed to determine whether screening improves mental health outcomes. Using the PHQ-2 in combination with the PHQ-9 may be a resource-efficient approach. Many individuals who screen positive, however, will not meet major depression diagnostic criteria and will need to be evaluated by a clinician.

Strengths of the study included the large sample size, inclusion of results from all cutoffs from all studies (rather than just those published), assessment of PHQ-2 accuracy separately across reference standards and by participant subgroups, and evaluation of the PHQ-2 and PHQ-9 combination, which had not been previously done in meta-analyses.

Limitations

This study has several limitations. First, primary data from 36 of 131 published eligible data sets (27\%) were not included. Second, there was moderate heterogeneity across studies, although it improved in most cases when subgroups were considered. Subgroup analyses based on medical comorbidities, as specified in the study protocol, and on country and language could not be conducted. This is because data on the presence of nonpsychiatric medical diagnoses were not available for 40\% of participants, with higher percentages missing for specific diagnoses, and because many countries and languages were represented in few primary studies.

Third, many included studies did not explicitly exclude participants who may have already been diagnosed or receiving care for depression, although there were not statistically significant differences between analyses of participants verified to not currently be diagnosed or receiving treatment and analyses of all participants, including those without this information.

Fourth, studies in the meta-analysis of individual participant data were categorized based on the interview administered, but it is possible that interviews may not have always been used in the way intended. Among 48 studies that used semistructured interviews, 3 used interviewers who did not meet typical standards, and 11 were rated unclear. It is possible that use of non-qualified interviewers may have reduced differences in accuracy estimates across reference standard categories.

Fifth, few studies were rated as having a low risk of bias across all QUADAS-2 domains; thus, sensitivity analyses using only studies with all low ratings were not conducted.

Conclusions

In an individual participant data meta-analysis of studies that compared PHQ scores with major depression diagnoses, the combination of PHQ-2 (with cutoff $\geq 2$) followed by PHQ-9 (with cutoff $\geq 10$) had similar sensitivity but higher specificity compared with PHQ-9 cutoff scores of 10 or greater alone. Further research is needed to understand the clinical and research value of this combined approach to screening.

\section*{Article Information}

\textbf{Accepted for Publication:} April 10, 2020.

\textbf{Author Affiliations:} Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada (Levis, Sun, Wu); Architectural and Occupational Health, McGill University, Montréal, Québec, Canada (Levis, Wu, Bhandari, Neupane, Negeri, Thombs); Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada (Levis, Wu, Bhandari, Neupane, Negeri, Thombs, Benedetti); Department of Psychiatry, McGill University, Montréal, Québec, Canada (Wu, Thombs); Center for Internal Medicine and Dermatology, Department of Psychosomatic Medicine, Charité, Universitätsmedizin Berlin, Germany (Fischer); Department of Medicine, McGill University, Montréal, Québec, Canada (Benedetti, Thombs); Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada (Benedetti); Department of Psychology, McGill University, Montréal, Québec, Canada (Thombs); Department of Educational and Counselling Psychology, McGill University, Montréal, Québec, Canada (Benedetti, Thombs).

\textbf{Author Contributions:} Drs Benedetti and Thombs had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Benedetti and Thombs contributed equally as co-senior authors. Concept and design: Levis, Benedetti, Thombs. Acquisition, analysis, or interpretation of data: All authors.

\textbf{Drafting of the manuscript:} Levis, Sun, Benedetti, Thombs.

\textbf{Critical review of the manuscript for important intellectual content:} All authors.

\textbf{Statistical analysis:} Levis, Sun, He, Wu, Negeri, Fischer, Benedetti, Thombs.

\textbf{Obtained funding:} Benedetti, Thombs.

\textbf{Administrative, technical, or material support:} Sun, Thombs.

\textbf{Supervision:} Benedetti, Thombs.

\textbf{Conflict of Interest Disclosures:} None reported.

\textbf{Funding/Support:} This study was funded by the Canadian Institutes of Health Research (CIHR; grants KRS-134297, PCG-155468, and PJT-162206). Dr Levis was supported by a CIHR Frederick Banting and Charles Best Canada Graduate Scholarship doctoral award and a Fonds de recherche du Québec-Santé (FRQS) Postdoctoral Training Fellowship. Ms Neupane was supported by a FRQS Postdoctoral Training Fellowship. Mr Bhandari was supported by a studentship from the Research Institute of the McGill University Health Centre. Ms Neupane was supported by G.R. Caverhill Fellowship from the Faculty of Medicine, McGill University. Drs Benedetti and Thombs were supported by FRQS researcher salary awards.

\textbf{Role of the Funder/Sponsor:} The funders had no 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.

\textbf{Group Information:} The DEPRESSD PHQ Collaboration members and contributions include the following:

\textbf{Data analysis:} Living Chen, McGill University, Montréal, Québec, Canada; and Alexander W. Levis, McGill University, Montréal, Québec, Canada.

\textbf{Data extraction, coding, and synthesis:} Kira E. Riehm, Lady Davis Institute for Medical Research, Montréal, Québec, Canada; and Lorie A. Kloda, Concordia University, Montréal, Québec, Canada.

\textbf{Design and conduct of database searches:} Jill Boruff, McGill University, Montréal, Québec, Canada; and Nazanin Saadat, Lady Davis Institute for Medical Research, Montréal, Québec, Canada.

\textbf{Supervision:} Pim Cuijpers, Vrije Universiteit, Amsterdam, the Netherlands; Simon Gilbody, University of York, Newcastle upon Tyne, UK; John P. A. Ioannidis, Stanford University, Stanford, California; Dean McMillan, University of York, Heslington, York, UK; John P. A. Ioannidis, Stanford University, Stanford, California; and Lorie A. Kloda, Concordia University, Montréal, Québec, Canada.

\textbf{DEPRESSD Steering Committee, including conception and oversight of collaboration:} Pim Cuijpers, Vrije Universiteit, Amsterdam, the Netherlands; Simon Gilbody, University of York, Heslington, York, UK; John P. A. Ioannidis, Stanford University, Stanford, California; Dean McMillan, University of York, Heslington, York, UK; Scott B. Patten, University of Calgary, Calgary, Alberta, Canada; Ian Shrier, McGill University, Montréal, Québec, Canada; and Roy C. Ziegelstein, Johns Hopkins University School of Medicine, Baltimore, Maryland.

\textbf{Knowledge user consultant:} Ainsley Moore, McMaster University, Hamilton, Ontario, Canada.

\textbf{Contributed included data sets:} Dickens H. Akena, Makerere University College of Health Sciences, Kampala, Uganda; Dagmar Amtmann, University of Washington, Seattle, Washington, USA; Bruce Arroll, University of Auckland, Auckland, New Zealand; List Ayalon, Bar Ilan University, Ramat Gan, Israel; Hamid R.® 2020 American Medical Association. All rights reserved.


