Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis

Brooke Levis,a,b Andrea Benedetti,b,c,d John P.A. Ioannisida,e,f,g,h Ying Suna, Zelalem Negeria,b, Chen Hea, Yin Wua,b,i, Ankur Krishnan,a, Parash Mani Bhandaria,b, Dipika Neupanea,b, Mahrukh Imranh, Danielle B. Ricen,j, Kira E. Riehm a,k, Nazanin Saadaht, Marleine Azara,b, Jill Boruffel, Pim Cuipjersm, Simon Gilboym, Lorie A. Kloado, Dean McMillank, Scott B. Pattenp,q, Ian Shrierabd, Roy C. Ziegelsteins, Sultan H. Alamrit, Dagmar Amtmannu, Liat Ayalonv, Hamid R. Baradaranw,x, Anna Beraldiy, Charles N. Bernsteinz, Arvin Bhana,b, Charles H. Bombardieru, Gregory Carterjj, Felix H. Fischeri,kk, Leila Gholizadehl, Lorna J. Gibsonmm, Eric P. Greennn, Catherine G. Greenooo, Brian J. Hallpp,qq, Emily E. Harozrr, Khalida Ismailss, Nathalie Jetttt, Mohammad E. Khamsehw, Yunxin Kwauu, Maria Asunción Laravv, Shen-Ing Liuw, Sonia R. Louereoee, Bernd Löwew, Ruth Ann Marriee, Laura Marshfff, Anthony McGuireddd, Kumiko Muramatsu, Laura Navarrettel, Flávia L. Osórió, Inge Petersenhh, Angelo Picardiiii, Stephanie L. Pughbbb, Terence J. Quinnll, Alasdair G. Rooney, Eileen H. Shinnnn, Abbey Sidebottomooo, Lena Spangenbergppp, Pei Lin Lynette Tanuu, Lynne I. Wagnernn, Jennifer Whitezzz, Kirsty Winkleyaaa, Brett D. Thombs*.a,b,c,d,i,j,kk

* Corresponding author. Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada.

https://doi.org/10.1016/j.jclinepi.2020.02.002

0895-4356/© 2020 Elsevier Inc. All rights reserved.
Abstract

Objectives: Depression symptom questionnaires are not for diagnostic classification. Patient Health Questionnaire-9 (PHQ-9) scores ≥10 are nonetheless often used to estimate depression prevalence. We compared PHQ-9 ≥10 prevalence to Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders (SCID) major depression prevalence and assessed whether an alternative PHQ-9 cutoff could more accurately estimate prevalence.

Study Design and Setting: Individual participant data meta-analysis of datasets comparing PHQ-9 scores to SCID major depression status.

Results: A total of 9,242 participants (1,389 SCID major depression cases) from 44 primary studies were included. Pooled PHQ-9 ≥10 prevalence was 24.6% (95% confidence interval [CI]: 20.8%, 28.9%); pooled SCID major depression prevalence was 12.1% (95% CI: 9.6%, 15.2%); and pooled difference was 11.9% (95% CI: 9.3%, 14.6%). The mean study-level PHQ-9 ≥10 to SCID-based prevalence ratio was 2.5 times. PHQ-9 ≥14 and the PHQ-9 diagnostic algorithm provided prevalence closest to SCID major depression prevalence, but study-level prevalence differed from SCID-based prevalence by an average absolute difference of 4.8% for PHQ-9 ≥10 to 14 and the PHQ-9 diagnostic algorithm provided prevalence closest to SCID major depression prevalence, but study-level prevalence differed from SCID-based prevalence by an average absolute difference of 4.8% for PHQ-9 ≥10 (95% prediction interval: −13.6%, 14.5%) and 5.6% for the PHQ-9 diagnostic algorithm (95% prediction interval: −16.4%, 15.0%).

Conclusion: PHQ-9 ≥10 substantially overestimates depression prevalence. There is too much heterogeneity to correct statistically in individual studies. © 2020 Elsevier Inc. All rights reserved.

Keywords: Depression prevalence; PHQ-9; SCID; Individual participant data meta-analysis

1. Introduction

Disease prevalence estimates have important implications for interpreting medical research, understanding disease burden, and making decisions about health care resource utilization [1]. In mental health research, major depression classification requires using validated diagnostic interviews [2,3]. Administering diagnostic interviews in large enough samples to estimate prevalence, however, is resource intensive. Thus, researchers sometimes use self-report depression symptom questionnaires, or screening tools, instead, and label the percentage of participants scoring above a screening cutoff as depression prevalence [4,5]. A 2018 study identified 19 primary studies listed in PubMed in a 3-month period whose titles indicated that they assessed the prevalence of depression or depressive disorders and found that 89% were based on screening questionnaires only [4].

Some self-report questionnaires include the same symptoms evaluated in validated diagnostic interviews. None, however, include all components of diagnostic interviews, such as assessment of functional impairment or investigation of nonpsychiatric medical conditions that can cause similar symptoms [4]. Using depression symptom questionnaires and cutoffs intended for screening to assess depression prevalence may overestimate prevalence. This is because screening attempts to identify previously unrecognized cases; cutoffs are set to cast a wide net and identify many more patients who may have depression than meet diagnostic criteria.

A recent review examined meta-analyses of depression prevalence published in 2008–2017 [5]. Of 81 prevalence estimates reported in abstracts of 69 meta-analyses, 10% were based on diagnostic interviews, 44% were based on screening or rating tools, and 46% combined results from diagnostic interviews and screening or rating tools. The mean reported prevalence was 31% among meta-analyses based on screening or rating tools compared with 17% with diagnostic interviews [5]. The degree to which screening tools exaggerate prevalence, however, depends on the screening tool and cutoff used [4,5].

We do not know of any studies that have evaluated the degree to which specific screening tool and cutoff combinations overestimate depression prevalence [4,5]. The Patient Health Questionnaire-9 (PHQ-9) [6–8] is the most commonly used depression screening tool in primary care [9]. Its nine items align with the Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria for major depressive episode (MDE) [10–12]. The standard cutoff, ≥10, is well established for screening to detect major depression and maximized combined sensitivity and specificity in a recent individual participant data meta-analysis (IPDMA) [6–8,13]. PHQ-9 ≥10 has been used to estimate depression prevalence in primary research studies and via synthesis in meta-analyses, including in very high-impact journals [14–16]. It is also sometimes used to diagnose depression and make treatment decisions for individual patients [6,17–19].

Our objective was to use an IPDMA approach to (1) compare PHQ-9 ≥10 prevalence to major depression prevalence based on a well-validated semistructured diagnostic interview, the Structured Clinical Interview for DSM (SCID) [20]; and (2) use a prevalence matching approach [4,21] to determine if a PHQ-9 cutoff could be set to match SCID-based prevalence with sufficiently low heterogeneity to accurately estimate prevalence in individual studies.

2. Methods

This study used a subset of data accrued for an IPDMA of the accuracy of the PHQ-9 for screening to detect major depression [13]. Detailed methods were registered in
PROSPERO (CRD42014010673), and a protocol was published [22]. This analysis was not part of the original IPD-MA protocol.

2.1. Study selection

In the main IPDMA, datasets from articles in any language were eligible for inclusion if (1) they included PHQ-9 scores; (2) they included diagnostic classifications for current MDE or major depressive disorder (MDD) based on DSM [10–12] or International Classification of Diseases [23] criteria, using a validated semistructured or fully structured interview; (3) the PHQ-9 and diagnostic interview were administered within 2 weeks of each other; (4) participants were aged ≥18 years and not recruited from youth or school-based settings; and (5) participants were not recruited from psychiatric settings or because they were identified as having depressive symptoms. Datasets where not all participants were eligible were included if primary data allowed the selection of eligible participants.

For the present study, we included primary studies that based diagnoses on the SCID [20]. The SCID is a semistructured diagnostic interview intended to be conducted by an experienced diagnostician; it requires clinical judgment and allows rephrasing questions and probes. The reason for including only SCID studies is that in analyses using large IPDMA databases [24–26], we found that, compared with semistructured interviews, fully structured interviews, which are designed for administration by lay interviewers, identify more participants with low-level symptoms as depressed but fewer participants with high-level symptoms. These results were consistent with the idea that semistructured interviews most closely replicate clinical interviews done by trained professionals, whereas fully structured interviews are less resource-intensive options that can be administered by research staff without diagnostic skills but may misclassify major depression in many participants. In our PHQ-9 IPDMA database, 44 of 47 studies that used semistructured interviews used the SCID. Thus, to reduce heterogeneity, we only included these 44 studies in main analyses.

In sensitivity analyses, we also included the three studies that used other semistructured interviews. We considered also incorporating published results from eligible studies that did not contribute data to the IPDMA. However, only 3 of 14 such studies [27–29] (970 participants, including 77 major depression cases) reported sufficient information to compare PHQ-9 ≥10 and SCID-based prevalence, and these studies did not report information necessary to be included in all prevalence matching analyses.

2.2. Data sources and searches

A medical librarian searched MEDLINE, MEDLINE In-Process & Other Non-Indexed Citations via Ovid, PsycINFO, and Web of Science from January 1, 2000, to May 9, 2018, using a peer-reviewed [30] search strategy (Supplementary Material: Appendix Methods). We also reviewed reference lists of relevant reviews and queried contributing authors about nonpublished studies.

Two investigators independently reviewed titles and abstracts for eligibility. If either deemed a study potentially eligible, the full text was reviewed by two investigators, independently, with disagreements resolved by consensus, consulting a third investigator when necessary.

2.3. Data contribution and synthesis

Authors of eligible datasets were invited to contribute deidentified primary data, including PHQ-9 scores and major depression classification status. We emailed corresponding authors of eligible studies at least three times, as necessary. If no response, we emailed coauthors and attempted phone contact.

What is new?

Key findings

- Based on 9,242 participants from 44 primary studies, the pooled Patient Health Questionnaire-9 (PHQ-9) ≥10 prevalence (25%) was double the pooled Structured Clinical Interview for DSM (SCID) major depression prevalence (12%); the pooled difference from individual studies was 12%.

- PHQ-9 ≥14 prevalence and PHQ-9 diagnostic algorithm prevalence most closely matched SCID major depression prevalence, but study-level PHQ-9 ≥14 and PHQ-9 diagnostic algorithm prevalence differed from SCID major depression prevalence with 95% prediction intervals of −14% to 15% and −16% to 15%, respectively.

What this adds to what was known?

- Although PHQ-9 ≥10 is often used to estimate depression prevalence, it overestimates major depression prevalence substantially.

- There is too much heterogeneity to correct statistically in individual studies.

What is the implication and what should change now?

- Estimates of depression prevalence should be based on validated diagnostic interviews designed for determining case status; users should evaluate published reports of depression prevalence to ensure that they are based on methods intended to classify major depression.
Before integrating individual datasets into our synthesized dataset, we compared published participant characteristics and diagnostic accuracy results with results from raw datasets and resolved discrepancies with the original investigators. When datasets included statistical weights to reflect sampling procedures, we used provided weights. For studies where sampling procedures merited weighting, but the original study did not weigh, we constructed weights using inverse selection probabilities.

2.4. Data analysis

2.4.1. Comparison of PHQ-9 ≥10 prevalence and SCID major depression prevalence

For each primary study, we estimated the percentage of participants who scored ≥10 on the PHQ-9, the percentage of participants classified as having major depression based on the SCID, the difference of these percentages, and the ratio. Then, across studies, we pooled prevalence for PHQ-9 ≥10, prevalence for the SCID, and differences in prevalence.

2.4.2. Prevalence matching

To identify which PHQ-9 scoring approach best matched SCID-based prevalence, we estimated pooled differences in prevalence for each possible PHQ-9 cutoff and the PHQ-9 diagnostic algorithm compared with SCID. The scoring approach with the smallest pooled difference was chosen to be the “prevalence match scoring approach.” Then, for each included study, we estimated the difference and ratio in prevalence for the prevalence match scoring approach vs. SCID. We determined the mean and median absolute difference and range of differences across all studies. To illustrate the range of difference values that would be expected if a new study were to compare prevalence based on the prevalence match scoring approach to prevalence based on SCID, we estimated 95% prediction intervals for the differences. For the diagnostic algorithm, which requires five or more items with scores of ≥2 points, with at least one being depressed mood or anhedonia [8], three studies [31–33] (524 participants) and 88 additional participants from other studies (612 participants total, 7%) were excluded, as they did not provide PHQ-9 item scores, which are necessary to determine diagnostic algorithm criteria. In sensitivity analyses, we evaluated if results differed if the 612 participants were excluded from all analyses rather than just those involving the diagnostic algorithm.

All meta-analyses incorporated sampling weights and were conducted in R (R version 3.4.1; R Studio version 1.0.143) using the lme4 package. To estimate pooled prevalence values, generalized linear mixed-effects models with a logit link function were fit using the glmer function. To estimate pooled difference values, linear mixed-effects models were fit using the lmer function. To account for the correlation between subjects within the same primary study, random intercepts were fit for each primary study. To quantify heterogeneity, we reported the estimated between-study variance (τ²) for each analysis.

In post-hoc analyses, we investigated whether differences in prevalence for the PHQ-9 prevalence match scoring approach and SCID were associated with study and participant characteristics. To do this, we fit additional linear mixed-effects models for pooled prevalence difference, including age, sex, country human development index (“very high,” “high,” or “low-medium,” based on the United Nation’s 2018 Human Development Index) and recruitment setting category (primary care, nonmedical care, inpatient specialty care, or outpatient specialty care) as fixed-effect covariates. For these analyses, we excluded 56 participants (<1%) missing age or sex data.

3. Results

3.1. Search results and inclusion of primary study datasets

Of 9,674 unique titles and abstracts identified from the database search for the main IPDMA, 9,198 were excluded after title and abstract review and 297 were excluded after full-text review, leaving 179 eligible articles with data from 123 unique participant samples, of which 95 (77.2%) contributed datasets. Authors of included studies contributed data from five unpublished studies, for a total of 100 datasets. Of these, for the present study’s main analyses, we excluded 56 studies that classified major depression using a diagnostic interview other than the SCID (Fig. 1). Thus, the main analyses of the present study included 9,242 participants (1,389 major depression cases) from 44 primary studies [31–72]. Among the 28 eligible primary studies that did not provide datasets for the main IPDMA, 14 used the SCID (4,408 participants). Thus, the main analyses included 75.9% of eligible studies that used the SCID (44/58) and 67.7% of eligible participants (9,242/13,650). Table 1 shows the characteristics of each included study.

In sensitivity analyses, we included data from three additional studies (1,992 participants, including 139 major depression cases) that provided individual participant data but administered a semistructured interview other than the SCID (Table 1) [73–75].

3.2. Comparison of PHQ-9 ≥10 prevalence and SCID major depression prevalence

The percentage of participants with PHQ-9 ≥10 in each of the 44 SCID studies ranged from 5.3% to 64.8%; pooled prevalence was 24.6% (95% confidence interval [CI]: 20.8%, 28.9%; τ²: 0.505). The percentage of participants with SCID major depression ranged from 0.6% to 56.4%; pooled prevalence was 12.1% (95% CI: 9.6%, 15.2%; τ²: 0.703).
Differences in prevalence (PHQ-9 ≥10 minus SCID) ranged from −6.0% to 46.9%. The pooled difference was 11.9% (95% CI: 9.3%, 14.6%; $\tau^2$: 0.007).

The ratio of PHQ-9 ≥10 prevalence to SCID-based prevalence ranged from 0.7 to 10.0 times (mean: 2.5 and median: 1.9). The mean ratio was 3.8 times for the 17 studies with SCID-based prevalence <10% (mean difference: 13.3%), 2.0 times for the 16 studies with SCID-based prevalence between 10% and 20% (mean difference: 12.7%), and 1.3 times for the 11 studies with SCID-based prevalence ≥20% (mean difference: 8.9%).

3.3. Prevalence matching

PHQ-9 ≥14 (pooled difference in prevalence: 0.5%, 95% CI: −1.7%, 2.6%, $\tau^2$: 0.005) and the PHQ-9 diagnostic algorithm (pooled difference in prevalence: −0.7%, 95% CI: −3.2%, 1.8%; $\tau^2$: 0.006) provided prevalence closest to SCID-based prevalence. Pooled differences in prevalence for PHQ-9 ≥13 and ≥15 compared with SCID were 2.6% and −2.0%.

In the 44 individual SCID studies, differences between the percentage of participants with PHQ-9 ≥14 and SCID major depression ranged from −18.7% to 29.7% (mean absolute difference: 4.8%). Of 44 prevalence estimates based on PHQ-9 ≥14, 24 (54.5%) were ≤0.75 times or ≥1.25 times the SCID-based prevalence. The 95% prediction interval for the difference in prevalence was −13.6% to 14.5%. For the PHQ-9 diagnostic algorithm, study-level differences in prevalence ranged from −20.1% to 27.1% (mean absolute difference: 5.6%). Of 41 prevalence estimates based on the PHQ-9 diagnostic algorithm, 28 (68.3%) were ≤0.75 times or ≥1.25 times the SCID-based prevalence. The 95% prediction interval for the
difference in prevalence was $-16.4\%$ to $15.0\%$. No study or participant characteristics were significantly associated with differences in prevalence for either of the PHQ-9 prevalence match scoring approaches compared with SCID.

3.4. Sensitivity analyses

Results for all analyses were similar when data from the three studies with semistructured interviews other than the SCID were added or when the 612 participants without data to determine PHQ-9 diagnostic algorithm classification were excluded.

4. Discussion

Primary studies and meta-analyses that describe their results as reflecting the prevalence of depression or depressive disorders are frequently based on depression screening tools, which are not designed for this purpose, rather than validated diagnostic interviews [4,5]. The PHQ-9 is often used to generate what are described by researchers as depression prevalence estimates. The present study found that using PHQ-9 $\geq 10$ to assess depression prevalence, which is commonly done, overestimated depression prevalence compared with prevalence based on actual diagnostic criteria by 11.9% (mean ratio: 2.5 times).

These results are consistent with what was predicted in a previous analysis that used hypothetical estimates of sensitivity and specificity to demonstrate how depression screening tools would be expected to inflate prevalence [4]. Results are also consistent with the findings of a meta-research review of prevalence estimates from 69 meta-analyses that found higher mean depression prevalence based on screening or rating tools than based on diagnostic interviews [5]. Thus, if a screening tool, such as the PHQ-9 $\geq 10$, is used to estimate prevalence, prevalence will appear to be substantial in virtually all populations, even when true prevalence is very low. This could have important ramifications in terms of policies, service planning, and health care budgets.

Identifying a PHQ-9 cutoff that could be used to match true prevalence based on a diagnostic interview would allow researchers to use inexpensive questionnaires instead of more costly interview methods for prevalence estimation. We tested a prevalence matching approach and found that PHQ-9 $\geq 14$ and the PHQ-9 diagnostic algorithm provided the smallest differences in prevalence compared with SCID major depression, but heterogeneity was high and not associated with study or participant characteristics. The mean absolute difference between prevalence based on PHQ-9 vs. SCID in individual studies was 4.8% for PHQ-9 $\geq 14$ and 6.3% for the PHQ-9 diagnostic algorithm, reflecting both overestimation and underestimation. For more than half of the studies examined, PHQ-9 $\geq 14$ prevalence was less than 75% or more than 125% of SCID-based prevalence; for the PHQ-9 diagnostic algorithm, the fraction was over two-thirds. The 95% prediction interval for the difference between PHQ-9 $\geq 14$ and SCID-based prevalence ranged from 14% below to 15% above SCID-based prevalence; for the PHQ-9 diagnostic algorithm, it was from 16% below to 15% above SCID-based prevalence.

Researchers sometimes report prevalence estimates based on cutoffs from questionnaires, including the PHQ-9, as prevalence of “clinically significant” symptoms or “symptoms” of depression, rather than “depression” [14,76,77]. However, screening tool cutoffs do not reflect a meaningful divide between impairment and non-impairment. Patients scoring at or above virtually any cutoff would be expected to have greater impairment than patients scoring below the cutoff, but no evidence has established any single cutoff for establishing an impairment threshold or that would support clinical decision-making for individual patients without a validated clinical assessment [4].

Research on screening using the PHQ-9 would be expected to report the proportion of patients who score at or above screening cutoffs because this provides information on the number of patients who would need resources for further mental health assessment. Reporting this percentage as depression prevalence, however, would be akin, for example, to reporting the proportion of women with positive mammogram screens as the prevalence of breast cancer and, as shown in the present study, would dramatically overestimate prevalence.

This is the first study to estimate the degree to which using PHQ-9 $\geq 10$ to estimate depression prevalence, a common practice, leads to overestimation of prevalence. The strengths of the study are that we incorporated data from 44 primary studies and that we directly compared PHQ-9 $\geq 10$ prevalence estimates to those based on the SCID, a rigorous semistructured interview intended to facilitate the standardized application of actual diagnostic criteria by trained diagnosticians [10–12]. This study had some limitations. First, we were unable to include data from 14 of 58 published eligible datasets (24%). Second, included datasets were almost exclusively from patients in health care settings where the presence of transdiagnostic somatic symptoms and adjustment to illness or injury may have contributed to error variance [75]. Third, included datasets were from a wide range of study settings, which may account for some of the observed heterogeneity. Fourth, the overestimation of prevalence when screening tools are used is expected to be greater with lower true prevalence. This is because false positives are disproportionately high in low-prevalence populations and only minimally offset by false negative screens, which occur when true cases are missed by the screening test. However, we were unable to assess this because of the small number of heterogeneous datasets included. Fifth, not all SCID studies described interviewer qualifications; untrained interviewers may have reduced the ability to detect differences across interviews. Sixth, we only examined one depression screening tool, the PHQ-9.
Studies from IPDMA that used other semistructured interviews and were included in sensitivity analyses participant missing data for age, sex, and diagnostic algorithm.

Studies from IPDMA that used the SCID and were included in main analyses

### Table 1. Characteristics of included studies and difference between percentage with PHQ-9 ≥10 and prevalence matching-based prevalence and prevalence based on diagnostic criteria for major depression

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Country</th>
<th>Recruited population</th>
<th>Total (N)</th>
<th>PHQ-9 ≥10</th>
<th>PHQ-9 ≥10, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alami, 2017*</td>
<td>Saudi Arabia</td>
<td>Hospitalized elderly in medical and surgical wards</td>
<td>199</td>
<td>68 (13)</td>
<td>67 (33.3)</td>
</tr>
<tr>
<td>Amoorez, 2017*</td>
<td>Canada</td>
<td>Migraine patients</td>
<td>203</td>
<td>102 (50)</td>
<td>62 (30.5)</td>
</tr>
<tr>
<td>Amtmann, 2015*</td>
<td>USA</td>
<td>Multiple sclerosis patients</td>
<td>164</td>
<td>55 (11)</td>
<td>171 (65.2)</td>
</tr>
<tr>
<td>Ayalon, 2010*</td>
<td>Israel</td>
<td>Elderly primary care patients</td>
<td>151</td>
<td>149 (76)</td>
<td>101 (66.8)</td>
</tr>
<tr>
<td>Beraldi, 2014*</td>
<td>Italy</td>
<td>Cancer inpatients</td>
<td>116</td>
<td>65 (16)</td>
<td>51 (16.1)</td>
</tr>
<tr>
<td>Berstein, 2018*</td>
<td>Canada</td>
<td>IBD patients</td>
<td>240</td>
<td>115 (16)</td>
<td>21 (8.8)</td>
</tr>
<tr>
<td>Bhana, 2015</td>
<td>South Africa</td>
<td>Chronic care patients</td>
<td>679</td>
<td>128 (17)</td>
<td>63 (9.2)</td>
</tr>
<tr>
<td>Bombardier, 2012</td>
<td>USA</td>
<td>Inpatients with spinal cord injuries</td>
<td>160</td>
<td>9 (6)</td>
<td>41 (25.6)</td>
</tr>
<tr>
<td>Chagas, 2013</td>
<td>Brazil</td>
<td>Outpatients with Parkinson’s Disease</td>
<td>84</td>
<td>59 (12)</td>
<td>39 (46)</td>
</tr>
<tr>
<td>Chibanda, 2016</td>
<td>Zimbabwe</td>
<td>A primary care population with high HIV prevalence</td>
<td>264</td>
<td>38 (10)</td>
<td>17 (6)</td>
</tr>
<tr>
<td>Eck, 2006</td>
<td>USA</td>
<td>Women seeking psychiatric services for their mental health at two mental health centers</td>
<td>404</td>
<td>81 (20)</td>
<td>13 (16)</td>
</tr>
<tr>
<td>Fann, 2005*</td>
<td>USA</td>
<td>Inpatients with traumatic brain injury</td>
<td>135</td>
<td>42 (31)</td>
<td>45 (33.3)</td>
</tr>
<tr>
<td>Fiest, 2014*</td>
<td>Canada</td>
<td>Epilepsy outpatients</td>
<td>169</td>
<td>150 (89)</td>
<td>125 (74)</td>
</tr>
<tr>
<td>Fischer, 2014*</td>
<td>Germany</td>
<td>Heart failure patients</td>
<td>194</td>
<td>105 (54)</td>
<td>44 (22.8)</td>
</tr>
<tr>
<td>Gjerdingen, 2009*</td>
<td>USA</td>
<td>Mothers registering their newborns for well-child visits at medical or pediatric clinics</td>
<td>419</td>
<td>30 (7)</td>
<td>19 (4.5)</td>
</tr>
<tr>
<td>Gräfe, 2004*</td>
<td>Germany</td>
<td>Medical and psychosomatic outpatients</td>
<td>494</td>
<td>31 (6)</td>
<td>46 (9.3)</td>
</tr>
<tr>
<td>Green, 2017</td>
<td>USA</td>
<td>Returning veterans</td>
<td>176</td>
<td>96 (54)</td>
<td>61 (35)</td>
</tr>
<tr>
<td>Green, 2018</td>
<td>Kenya</td>
<td>Pregnant women and new mothers</td>
<td>192</td>
<td>66 (34)</td>
<td>11 (6)</td>
</tr>
<tr>
<td>Haroz, 2017</td>
<td>Myanmar</td>
<td>Primary care patients</td>
<td>132</td>
<td>86 (65)</td>
<td>29 (22)</td>
</tr>
<tr>
<td>Hitchon, 2019*</td>
<td>Canada</td>
<td>Rheumatoid arthritis patients</td>
<td>148</td>
<td>124 (84)</td>
<td>16 (11)</td>
</tr>
<tr>
<td>Khamseh, 2011*</td>
<td>Iran</td>
<td>Type 2 diabetes patients</td>
<td>184</td>
<td>96 (52)</td>
<td>79 (42.9)</td>
</tr>
<tr>
<td>Kwan, 2012</td>
<td>Singapore</td>
<td>Post-stroke inpatients undergoing rehabilitation</td>
<td>113</td>
<td>37 (33)</td>
<td>3 (2.7)</td>
</tr>
<tr>
<td>Lambert, 2015</td>
<td>Australia</td>
<td>Cancer patients</td>
<td>147</td>
<td>96 (65)</td>
<td>21 (14.3)</td>
</tr>
<tr>
<td>Lara, 2015</td>
<td>Mexico</td>
<td>Pregnant women during the third trimester of pregnancy</td>
<td>280</td>
<td>280 (100)</td>
<td>29 (10.4)</td>
</tr>
<tr>
<td>Marrie, 2015</td>
<td>Canada</td>
<td>Multiple sclerosis patients</td>
<td>244</td>
<td>198 (81)</td>
<td>25 (10.2)</td>
</tr>
<tr>
<td>Martin-Subero, 2017</td>
<td>Spain</td>
<td>Medical inpatients</td>
<td>1,003</td>
<td>457 (45)</td>
<td>83 (8.3)</td>
</tr>
<tr>
<td>Osório, 2009</td>
<td>Brazil</td>
<td>Women in primary care</td>
<td>177</td>
<td>171 (100)</td>
<td>60 (35.9)</td>
</tr>
<tr>
<td>Osório, 2012</td>
<td>Brazil</td>
<td>Inpatients from various clinical wards</td>
<td>86</td>
<td>35 (41)</td>
<td>28 (32.6)</td>
</tr>
<tr>
<td>Patten, 2015</td>
<td>Canada</td>
<td>Multiple sclerosis patients</td>
<td>143</td>
<td>110 (77)</td>
<td>20 (14.0)</td>
</tr>
<tr>
<td>Picard, 2005</td>
<td>Italy</td>
<td>Inpatients with skin diseases</td>
<td>138</td>
<td>77 (56)</td>
<td>12 (8.7)</td>
</tr>
<tr>
<td>Prisco, 2016</td>
<td>Canada</td>
<td>Stroke and transient ischemic attack patients</td>
<td>114</td>
<td>64 (56)</td>
<td>11 (9.6)</td>
</tr>
<tr>
<td>Quinn, unpublished</td>
<td>UK</td>
<td>Stroke patients</td>
<td>146</td>
<td>65 (47)</td>
<td>17 (11.6)</td>
</tr>
<tr>
<td>Richardson, 2010</td>
<td>Canada</td>
<td>Older adults undergoing in-home aging services care management assessment</td>
<td>377</td>
<td>258 (68)</td>
<td>95 (25.2)</td>
</tr>
<tr>
<td>Rooney, 2013</td>
<td>UK</td>
<td>Adults with cervicalgia</td>
<td>126</td>
<td>54 (43)</td>
<td>14 (11.1)</td>
</tr>
<tr>
<td>Shinn, 2017*</td>
<td>USA</td>
<td>Cancer patients</td>
<td>139</td>
<td>109 (80)</td>
<td>10 (8.6)</td>
</tr>
<tr>
<td>Sidebottom, 2012*</td>
<td>USA</td>
<td>Pregnant women</td>
<td>246</td>
<td>246 (100)</td>
<td>12 (4.9)</td>
</tr>
<tr>
<td>Simring, 2012*</td>
<td>USA</td>
<td>Older adults living in public housing</td>
<td>190</td>
<td>110 (58)</td>
<td>10 (5.3)</td>
</tr>
<tr>
<td>Spangenberg, 2015</td>
<td>Germany</td>
<td>Primary care patients</td>
<td>160</td>
<td>97 (61)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td>Turner, 2012</td>
<td>Australia</td>
<td>Stroke patients</td>
<td>72</td>
<td>45 (37)</td>
<td>13 (18.1)</td>
</tr>
<tr>
<td>Turner, unpublished</td>
<td>Australia</td>
<td>Cardiovascular rehabilitation patients</td>
<td>51</td>
<td>7 (14)</td>
<td>4 (7.8)</td>
</tr>
<tr>
<td>Vöhringer, 2013*</td>
<td>Chile</td>
<td>Primary care patients</td>
<td>190</td>
<td>143 (75)</td>
<td>59 (31.1)</td>
</tr>
<tr>
<td>Wagner, 2017*</td>
<td>USA</td>
<td>Patients starting radiotherapy for the first diagnosis of any tumor</td>
<td>54</td>
<td>38 (69)</td>
<td>6 (4.3)</td>
</tr>
<tr>
<td>Williams, 2012</td>
<td>USA</td>
<td>Parkinson’s Disease patients</td>
<td>235</td>
<td>76 (32)</td>
<td>61 (26.0)</td>
</tr>
<tr>
<td>Wittkamp, 2009*</td>
<td>Netherlands</td>
<td>Primary care patients at risk for depression</td>
<td>260</td>
<td>175 (64)</td>
<td>45 (17.6)</td>
</tr>
</tbody>
</table>

Studies from IPDMA that used other semistructured interviews and were included in sensitivity analyses

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Country</th>
<th>Recruited population</th>
<th>Total (N)</th>
<th>PHQ-9 ≥10</th>
<th>PHQ-9 ≥10, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu, 2011*</td>
<td>Taiwan</td>
<td>Primary care patients</td>
<td>1,532</td>
<td>933 (61)</td>
<td>50 (3.3)</td>
</tr>
<tr>
<td>McGuire, 2013*</td>
<td>USA</td>
<td>Acute coronary syndrome inpatients</td>
<td>100</td>
<td>31 (31)</td>
<td>9 (9.0)</td>
</tr>
<tr>
<td>Twist, 2013*</td>
<td>UK</td>
<td>Type 2 diabetes outpatients</td>
<td>360</td>
<td>172 (45)</td>
<td>80 (7.4)</td>
</tr>
</tbody>
</table>

Abbreviations: DA-i, positive classification based on the PHQ-9 diagnostic algorithm.

* Study did not provide item-level data necessary to determine classification based on the PHQ-9 diagnostic algorithm.

+ Sampling weights were applied. Counts are based on actual numbers, whereas percentages are weighted.

# One participant missing data for age.

+ Ten participants missing data for age.

# One participant missing data for both age and sex.

+ Two participants missing data for age.

* Twenty-one participants missing data for age.

+ Unpublished at the time of electronic database search.

+ Sixty-two participants missing data to determine classification based on the PHQ-9 diagnostic algorithm.

+ Two participants missing data for age, seven participants missing data for sex, 10 participants missing data to determine classification based on the PHQ-9 diagnostic algorithm, and one participant missing data for age, sex, and diagnostic algorithm.

+ Two participants missing data for age, four participants missing data to determine classification based on the PHQ-9 diagnostic algorithm, and one participant missing data for age and diagnostic algorithm.

+ Four participants missing data for age.

+ Diagnostic interview: Schedules for Clinical Assessment in Neuropsychiatry.

+ Diagnostic interview: Depression Interview and Structured Hamilton.

+ Eight participants missing data to determine classification based on the PHQ-9 diagnostic algorithm.
### PHQ-9 

#### Prevalence matching

<table>
<thead>
<tr>
<th>% difference: PHQ-9 ≥10</th>
<th>Ratio: PHQ-9 ≥10/Major depression</th>
<th>% difference: PHQ-9 ≥14</th>
<th>Ratio: PHQ-9 ≥14/Major depression</th>
<th>% difference: PHQ-9 DA+</th>
<th>Ratio: PHQ-9 DA+/Major depression</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ-9 ≥10, n (%)</td>
<td>PHQ-9 ≥14, n (%)</td>
<td>PHQ-9 DA+, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.1</td>
<td>1.8</td>
<td>25 (12.6)</td>
<td>0.5</td>
<td>1.0</td>
<td>—</td>
</tr>
<tr>
<td>11.3</td>
<td>1.5</td>
<td>40 (19.7)</td>
<td>−4.4</td>
<td>0.8</td>
<td>36 (17.7)</td>
</tr>
<tr>
<td>15.4</td>
<td>1.9</td>
<td>55 (20.2)</td>
<td>2.6</td>
<td>1.1</td>
<td>42 (15.4)</td>
</tr>
<tr>
<td>5.3</td>
<td>2.3</td>
<td>7 (4.6)</td>
<td>0.7</td>
<td>1.2</td>
<td>6 (4.0)</td>
</tr>
<tr>
<td>12.1</td>
<td>3.0</td>
<td>4 (3.4)</td>
<td>−2.6</td>
<td>0.6</td>
<td>2 (1.7)</td>
</tr>
<tr>
<td>15.8</td>
<td>2.8</td>
<td>33 (13.8)</td>
<td>5.0</td>
<td>1.6</td>
<td>25 (10.4)</td>
</tr>
<tr>
<td>3.7</td>
<td>0.7</td>
<td>26 (3.8)</td>
<td>−7.7</td>
<td>0.3</td>
<td>15 (2.2)</td>
</tr>
<tr>
<td>18.1</td>
<td>3.1</td>
<td>23 (14.4)</td>
<td>5.6</td>
<td>1.6</td>
<td>17 (10.6)</td>
</tr>
<tr>
<td>13.1</td>
<td>1.6</td>
<td>16 (19.0)</td>
<td>−3.6</td>
<td>0.8</td>
<td>12 (14.3)</td>
</tr>
<tr>
<td>8.3</td>
<td>1.1</td>
<td>122 (46.2)</td>
<td>−10.2</td>
<td>0.8</td>
<td>96 (36.4)</td>
</tr>
<tr>
<td>25.0</td>
<td>2.0</td>
<td>17 (35.4)</td>
<td>10.4</td>
<td>1.4</td>
<td>17 (35.4)</td>
</tr>
<tr>
<td>6.3</td>
<td>1.4</td>
<td>33 (12.2)</td>
<td>−4.0</td>
<td>0.8</td>
<td>—</td>
</tr>
<tr>
<td>8.3</td>
<td>1.6</td>
<td>17 (10.1)</td>
<td>−3.6</td>
<td>0.7</td>
<td>17 (10.1)</td>
</tr>
<tr>
<td>13.4</td>
<td>3.4</td>
<td>19 (9.8)</td>
<td>4.1</td>
<td>1.7</td>
<td>20 (10.3)</td>
</tr>
<tr>
<td>7.2</td>
<td>2.6</td>
<td>26 (6.2)</td>
<td>1.7</td>
<td>1.4</td>
<td>31 (7.4)</td>
</tr>
<tr>
<td>20.0</td>
<td>2.5</td>
<td>97 (19.6)</td>
<td>6.1</td>
<td>1.4</td>
<td>86 (17.4)</td>
</tr>
<tr>
<td>24.4</td>
<td>3.0</td>
<td>31 (17.6)</td>
<td>5.1</td>
<td>1.4</td>
<td>32 (18.2)</td>
</tr>
<tr>
<td>46.9</td>
<td>10.0</td>
<td>67 (34.9)</td>
<td>29.7</td>
<td>6.7</td>
<td>62 (32.3)</td>
</tr>
<tr>
<td>−3.0</td>
<td>0.9</td>
<td>16 (12.1)</td>
<td>−9.8</td>
<td>0.6</td>
<td>13 (9.8)</td>
</tr>
<tr>
<td>18.9</td>
<td>2.8</td>
<td>22 (14.9)</td>
<td>4.1</td>
<td>1.4</td>
<td>26 (17.6)</td>
</tr>
<tr>
<td>13.0</td>
<td>1.3</td>
<td>81 (44.0)</td>
<td>1.1</td>
<td>1.0</td>
<td>55 (45.1)</td>
</tr>
<tr>
<td>18.6</td>
<td>8.0</td>
<td>9 (8.0)</td>
<td>5.3</td>
<td>3.0</td>
<td>7 (6.2)</td>
</tr>
<tr>
<td>11.6</td>
<td>1.6</td>
<td>21 (14.3)</td>
<td>0.0</td>
<td>1.0</td>
<td>18 (12.2)</td>
</tr>
<tr>
<td>10.0</td>
<td>2.0</td>
<td>21 (7.5)</td>
<td>−2.9</td>
<td>0.7</td>
<td>23 (8.2)</td>
</tr>
<tr>
<td>19.7</td>
<td>2.9</td>
<td>43 (17.6)</td>
<td>7.4</td>
<td>1.7</td>
<td>36 (14.8)</td>
</tr>
<tr>
<td>20.5</td>
<td>3.5</td>
<td>154 (15.4)</td>
<td>7.1</td>
<td>1.9</td>
<td>143 (14.3)</td>
</tr>
<tr>
<td>11.1</td>
<td>1.0</td>
<td>45 (25.4)</td>
<td>−8.5</td>
<td>0.8</td>
<td>43 (24.3)</td>
</tr>
<tr>
<td>15.1</td>
<td>1.5</td>
<td>26 (30.2)</td>
<td>−2.3</td>
<td>0.9</td>
<td>26 (30.2)</td>
</tr>
<tr>
<td>11.2</td>
<td>1.8</td>
<td>24 (16.8)</td>
<td>2.8</td>
<td>1.2</td>
<td>12 (8.4)</td>
</tr>
<tr>
<td>18.8</td>
<td>3.2</td>
<td>21 (15.2)</td>
<td>6.5</td>
<td>1.8</td>
<td>18 (13.0)</td>
</tr>
<tr>
<td>4.4</td>
<td>1.5</td>
<td>11 (9.6)</td>
<td>0.0</td>
<td>1.0</td>
<td>9 (7.9)</td>
</tr>
<tr>
<td>17.8</td>
<td>2.5</td>
<td>17 (11.6)</td>
<td>0.5</td>
<td>1.0</td>
<td>17 (12.6)</td>
</tr>
<tr>
<td>5.8</td>
<td>1.2</td>
<td>65 (17.2)</td>
<td>−8.0</td>
<td>0.7</td>
<td>60 (15.9)</td>
</tr>
<tr>
<td>10.3</td>
<td>1.9</td>
<td>15 (11.9)</td>
<td>0.8</td>
<td>1.1</td>
<td>13 (10.3)</td>
</tr>
<tr>
<td>8.6</td>
<td>2.0</td>
<td>11 (7.9)</td>
<td>−0.7</td>
<td>0.9</td>
<td>8 (6.5)</td>
</tr>
<tr>
<td>19.1</td>
<td>4.9</td>
<td>32 (13.0)</td>
<td>8.1</td>
<td>2.7</td>
<td>32 (13.0)</td>
</tr>
<tr>
<td>7.9</td>
<td>2.5</td>
<td>11 (5.8)</td>
<td>0.5</td>
<td>1.1</td>
<td>9 (4.7)</td>
</tr>
<tr>
<td>5.0</td>
<td>9.0</td>
<td>4 (2.5)</td>
<td>1.9</td>
<td>4.0</td>
<td>4 (2.5)</td>
</tr>
<tr>
<td>12.5</td>
<td>1.7</td>
<td>12 (16.7)</td>
<td>−1.4</td>
<td>0.9</td>
<td>9 (12.5)</td>
</tr>
<tr>
<td>3.9</td>
<td>1.5</td>
<td>2 (3.9)</td>
<td>−3.9</td>
<td>0.5</td>
<td>2 (3.9)</td>
</tr>
<tr>
<td>13.7</td>
<td>1.4</td>
<td>54 (28.4)</td>
<td>−2.6</td>
<td>0.9</td>
<td>—</td>
</tr>
<tr>
<td>0.9</td>
<td>1.2</td>
<td>7 (2.8)</td>
<td>−1.5</td>
<td>0.7</td>
<td>6 (2.4)</td>
</tr>
<tr>
<td>−6.0</td>
<td>0.8</td>
<td>17 (7.2)</td>
<td>−18.7</td>
<td>0.3</td>
<td>17 (7.2)</td>
</tr>
<tr>
<td>10.6</td>
<td>1.9</td>
<td>49 (11.6)</td>
<td>0.0</td>
<td>1.0</td>
<td>44 (10.4)</td>
</tr>
</tbody>
</table>

Studies from IPDMA that used the SCID and were included in main analyses

<table>
<thead>
<tr>
<th>% difference: PHQ-9 DA+</th>
<th>Ratio: PHQ-9 DA+/Major depression</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ-9 DA+, n (%)</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>2.7</td>
</tr>
<tr>
<td>16.0</td>
<td>2.8</td>
</tr>
<tr>
<td>7.4</td>
<td>2.0</td>
</tr>
</tbody>
</table>
although we expect that other tools would similarly exaggerate depression prevalence [4,5].

In summary, we found that using PHQ-9 ≥10 to estimate depression prevalence results in estimates that are, on average, 12% greater than what would be obtained using validated semistructured diagnostic interviews. Substantial heterogeneity presents a barrier to using statistical methods to estimate major depression prevalence based on PHQ-9 ≥10 or based on any other PHQ-9 cutoff. Researchers should not report results from the PHQ-9 as prevalence of major depression. Users of evidence should evaluate reports of prevalence with caution and ensure that they are based on methods intended to classify major depression.

CRediT authorship contribution statement

Brooke Levis: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. Andrea Benedetti: Conceptualization, Methodology, Formal analysis, Writing - review & editing. John P.A. Ioannidis: Conceptualization, Methodology, Writing - review & editing. Ying Sun: Formal analysis, Writing - review & editing. Chen He: Formal analysis, Writing - review & editing. Yan Wu: Formal analysis, Writing - review & editing. Ankur Krishnan: Formal analysis, Writing - review & editing. Parash Mani Bhandari: Formal analysis, Writing - review & editing. Dipika Neupane: Formal analysis, Writing - review & editing. Mahrukh Imran: Formal analysis, Writing - review & editing. Danielle B. Rice: Formal analysis, Writing - review & editing. Nazanin Saadat: Formal analysis, Writing - review & editing. Marline Azar: Formal analysis, Writing - review & editing. Jill Boruff: Conceptualization, Methodology, Investigation, Writing - review & editing. Pim Cuijpers: Conceptualization, Methodology, Writing - review & editing. Simon Gilbody: Conceptualization, Methodology, Writing - review & editing. Lorie A. Kloda: Conceptualization, Methodology, Investigation, Writing - review & editing. Dean McMillan: Conceptualization, Methodology, Writing - review & editing. Scott B. Patten: Conceptualization, Methodology, Data curation, Writing - review & editing. Ian Shrier: Conceptualization, Methodology, Writing - review & editing. Roy C. Ziegelstein: Conceptualization, Methodology, Writing - review & editing. Sultan H. Alamri: Data curation, Writing - review & editing. Dagmar Amtmann: Data curation, Writing - review & editing. Liat Ayalon: Data curation, Writing - review & editing. Hamid R. Baradaran: Data curation, Writing - review & editing. Anna Beraldi: Data curation, Writing - review & editing. Charles N. Bernstein: Data curation, Writing - review & editing. Arvin Bhana: Data curation, Writing - review & editing. Charles H. Bombardier: Data curation, Writing - review & editing. Gregory Carter: Data curation, Writing - review & editing. Marcos H. Chagas: Data curation, Writing - review & editing. Dixon Chibanda: Data curation, Writing - review & editing. Kerrie Clover: Data curation, Writing - review & editing. Yeates Conwell: Data curation, Writing - review & editing. Crisanto Díez-Quevedo: Data curation, Writing - review & editing. Jesse R. Fann: Data curation, Writing - review & editing. Felix H. Fischer: Data curation, Writing - review & editing. Leila Gholizadeh: Data curation, Writing - review & editing. Lorna J. Gibson: Data curation, Writing - review & editing. Eric P. Green: Data curation, Writing - review & editing. Catherine G. Greeno: Data curation, Writing - review & editing. Brian J. Hall: Data curation, Writing - review & editing. Emily E. Haroz: Data curation, Writing - review & editing. Khalida Ismail: Data curation, Writing - review & editing. Mohammad E. Khamehchi: Data curation, Writing - review & editing. Yunxin Kwan: Data curation, Writing - review & editing. Maria Asunción Lara: Data curation, Writing - review & editing. Shen-Ing Liu: Data curation, Writing - review & editing. Sonia R. Loureiro: Data curation, Writing - review & editing. Bernd Lowe: Data curation, Writing - review & editing. Ruth Ann Marrie: Data curation, Writing - review & editing. Laura Marsh: Data curation, Writing - review & editing. Anthony McGuire: Data curation, Writing - review & editing. Kumiko Muramatsu: Data curation, Writing - review & editing. Laura Navarrete: Data curation, Writing - review & editing. Flávia L. Osório: Data curation, Writing - review & editing. Inge Petersen: Data curation, Writing - review & editing. Angelo Picardi: Data curation, Writing - review & editing. Stephanie L. Pugh: Data curation, Writing - review & editing. Terence J. Quinn: Data curation, Writing - review & editing. Alasdair G. Rooney: Data curation, Writing - review & editing. Eileen H. Shinn: Data curation, Writing - review & editing. Abbey Sidebottom: Data curation, Writing - review & editing. Lena Spangenberg: Data curation, Writing - review & editing. Pei Lin Lynnette Tan: Data curation, Writing - review & editing. Martin Taylor-Rowan: Data curation, Writing - review & editing. Alyna Turner: Data curation, Writing - review & editing. Henk C. van Weert: Data curation, Writing - review & editing. Paul A. Vöhringer: Data curation, Writing - review & editing. Lynne I. Wagner: Data curation, Writing - review & editing. Jennifer White: Data curation, Writing - review & editing. Kirsty Winkley: Data curation, Writing - review & editing. Brett D. Thoms: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing.

Acknowledgments


Funding: This study was funded by the Canadian Institutes of Health Research (CIHR; KRS-134297, PCG-155468, PJT-162206). Dr. Levis and Dr. Wu were supported by Fonds de recherche du Québec - Santé (FRQs) Postdoctoral Training Fellowships. Drs. Benedetti and Thombs were supported by an FRQs researcher salary award. 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. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. Ms. Riehm and Ms. Saadat were supported by CIHR Frederick Banting and Charles Best Canada Graduate Scholarship master’s awards. The primary studies by Fiest et al., Patten et al., Amoozegar et al. and Prisnie et al. were supported by the Cumming School of Medicine, University of Calgary, and Alberta Health Services through the Calgary Health Trust, as well as the Hotchkiss Brain Institute. Dr. Patten was supported by a Senior Health Scholar award from Alberta Innovates Health Solutions. Dr. Jetté was supported by a Canada Research Chair in Neurological Health Services Research and an AIHS Population Health Investigator Award. The primary study by Amtmann et al. was supported by a grant from the Department of Education (NIDRR grant number H133B080025) and by the National Multiple Sclerosis Society (MB 0008). Data collection for the study by Ayalon et al. was supported from a grant from Lundbeck International. The primary study by Khamsheh et al. was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary studies by Marrie et al. and Bernstein et al. were supported by CIHR (THC-135234) and Crohn’s and Colitis Canada. Dr. Bernstein was supported in part by the Bingham Chair in Gastroenterology. Dr. Marrie was supported by the Waugh Family Chair in Multiple Sclerosis. The primary study by Bhana et al. was output of the Program for Improving Mental health careE (PRIME) and was supported by the UK Department for International Development (201446). The views expressed do not necessarily reflect the UK Government’s official policies. The primary study by Bombardier et al. was supported by the Department of Education, National Institute on Disability and Rehabilitation Research, Spinal Cord Injury Model Systems: University of Washington (grant no. H133N060033), Baylor College of Medicine (grant no. H133M060003), and University of Michigan (grant no. H133M060032). The primary study by Chibanda et al. was supported by a grant from Grand Challenges Canada (0087-04). Dr. Conwell received support from NIMH (R24MH071604) and the Centers for Disease Control and Prevention (R49 CE002093). The primary study by Martin-Subero et al. was supported in part by a grant from the Spanish Ministry of Health’s Health Research Fund (Fondo de Investigaciones Sanitarias, project 97/1184). Collection of data for the primary study by Fann et al. was supported by grant RO1 HD39415 from the US National Center for Medical Rehabilitation Research. The primary study by Fischer et al. was funded by the German Federal Ministry of Education and Research (01GY1150). Collection of data for the primary study by Gjerdingen et al. was supported by grants from the NIMH (R34 MH072925, K02 MH65919, P30 DK50456). The primary study by Gholizadeh et al. was supported by University of Technology Sydney under UTS Research Reestablishment Grants. The primary study by Green et al. (2018) was supported by a grant from the Duke Global Health Institute (453-0751). The primary study by Eack et al. was funded by the NIMH (R24 MH56858). The primary study by Garabiles et al. was supported by the Macao SAR Government, through the University of Macau RSKTO grants: MYRG-2014-111. The primary study by Haroz et al. was supported by the United States Agency for International Development Victims of Torture Fund: AID-DFD A-00-08-00308. Dr. Haroz was supported by a NIMH T32 predoctoral training grant (MH014592-38) and postdoctoral training grant (MH103210) during the conduct of primary study. 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 number RP-PG-0606-1142). The primary study by Lara et al. was supported by the Consejo Nacional de Ciencia y Tecnología/ National Council for Science and Technology (CB-2009-133923-H). The primary study by Liu et al. (2011) was funded by a grant from the National Health Research Institute, Republic of China (NHRI-EX97-9706PI). The primary studies by Osório et al. (2012) were funded by the German Federal Ministry of Education and Research under its Programme Grants for Applied Research Programme (grant number 09.1.01689.17.7) and Banco Santander (grant number 10.1.01232.17.9). Dr. Bernd Löwe received research grants from Pfizer, Germany, and from the medical faculty of the University of Heidelberg, Germany (project 121/2000) for the study by Gräfe et al. Dr. Marrie was supported
by the Waugh Family Chair in Multiple Sclerosis and the Research Manitoba Chair, and CIHR grants, during the conduct of the study. Collection of data for the primary study by Williams et al. was supported by a NIMH grant to Dr. Marsh (RO1-MH069666). The primary study by Muramatsu et al. (2018) was supported by grants from Nigata Seiryo University. Dr. Osorio was supported by Productivity Grants (PQ-CNpq-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. The primary study by Wagner et al. was supported by grants U10CA21661, U10CA180868, U10CA180822, and U10CA37422 from the National Cancer Institute. The study was also funded in part by a grant from the Pennsylvania Department of Health. The department specifically declares responsibility for any analyses, interpretations, or conclusions of the primary study. The primary study by Rooney et al. was funded by the United Kingdom National Health Service Lothian Neuro-Oncology Endowment Fund. The primary study by Shinn et al. was supported by grant NCI K07 CA 093512 and the Lance Armstrong Foundation. 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 number R44MC07840). Sinning et al. research was supported in part by grants from the NIH (T32 GM07356), Agency for Healthcare Research and Quality (R36 HS018246), NIMH (R24 MH017604), and the National Center for Research Resources (TL1 RR024135). The primary study by Spangenber et al. was supported by a junior research grant from the medical faculty, University of Leipzig. Collection of data for the studies by Turner et al. (2012) were funded by a bequest from Jennie Thomas through the Hunter Medical Research Institute. The study by Wittkamp et al. was funded by the Netherlands Organization for Health Research and Development (ZonMw) Mental Health Program (nos. 100.003.005 and 100.002.021) and the Academic Medical Center/University of Amsterdam. Dr Vöhringer was supported by the Fund for Innovation and Competitiveness of the Chilean Ministry of Economy, Development and Tourism, through the Millennium Scientific Initiative (grant number IS130005). No other authors reported funding for primary studies or for their work on the present study. No sponsor or funder was involved in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Conflicts of interests: All authors have completed the ICJME uniform disclosure form and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years with the following exceptions: C.N.B. declares that he receives grants and personal fees from Abbvie, Janssen, Pfizer, and Takeda; grants from Shire Canada, Celgene, Boeringher Ingelheim, and Roche; and personal fees from Mylan Pharmaceuticals; outside the submitted work. K.I. declares that she has received an honorarium for speaker fees for educational lectures for Sanofi, Sunovion, Janssen, and Novo Nordisk. S.L.P. declares that she received salary support from Pfizer-Astella and Millennium, outside the submitted work. L.I.W. declares that she receives personal fees from Celgene, outside the submitted work. All authors declare no other relationships or activities that could appear to have influenced the submitted work. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data sharing: Statistical codes and dataset used in the individual patient data meta-analysis can be requested from the corresponding author, Dr. Brett D. Thombs.

References


[52] Khamseh ME, Baradaran HR, Javanbakht A, Mirghobadian M, Yadollahi Z, Malek M. Comparison of the CES-D and PHQ-9

Kwan Y, Tham WY, Ang A. Validity of the Patient Health Questionnaire-9 (PHQ-9) in the screening of post-stroke depression in a multi-ethnic population. Biol Psychiatry 2012;71:141S.


Appendix Methods. Search Strategies

MEDLINE (OvidSP)
1. PHQ*.af.
2. patient health questionnaire*.af.
3. 1 or 2.
4. Mass Screening/.
5. Psychiatric Status Rating Scales/.
6. "Predictive Value of Tests"/.
7. "Reproducibility of Results"/.
8. exp "Sensitivity and Specificity"/.
9. Psychometrics/.
11. Reference Values/.
12. Reference Standards/.
13. exp Diagnostic Errors/.
14. Mental Disorders/di, pc [Diagnosis, Prevention & Control].
15. Mood Disorders/di, pc [Diagnosis, Prevention & Control].
16. Depressive Disorder/di, pc [Diagnosis, Prevention & Control].
17. Depressive Disorder, Major/di, pc [Diagnosis, Prevention & Control].
18. Depression, Postpartum/di, pc [Diagnosis, Prevention & Control].
19. Depression/di, pc [Diagnosis, Prevention & Control].
20. validation studies.pt.
22. screen*.af.
23. prevalence.af.
24. predictive value*.af.
25. detect*.ti.
26. sensitiv*.ti.
27. valid*.ti.
28. revalid*.ti.
29. predict*.ti.
30. accura*.ti.
31. psychometric*.ti.
32. identif*.ti.
33. specificit*.ab.
34. cut?off*.ab.
35. cut* score*.ab.
36. cut?point*.ab.
37. threshold score*.ab.
38. reference standard*.ab.
39. reference test*.ab.
40. index test*.ab.
41. gold standard.ab.
42. or/4-41.
43. 3 and 42.
44. limit 43 to yr = "2000-Current".

PsycINFO (OvidSP)
1. PHQ*.af.
2. patient health questionnaire*.af.
3. 1 or 2.
4. Diagnosis/.
5. Medical Diagnosis/.
6. Psychodiagnosis/.
7. Misdiagnosis/.
8. Screening/.
9. Health Screening/.
10. Screening Tests/.
11. Prediction/.
12. Cutting Scores/.
13. Psychometrics/.
14. Test Validity/.
15. screen*.af.
16. predictive value*.af.
17. detect*.ti.
18. sensitiv*.ti.
19. valid*.ti.
20. revalid*.ti.
22. psychometric*.ti.
23. specificit*.ab.
24. cut?off*.ab.
25. cut* score*.ab.
27. threshold score*.ab.
28. reference standard*.ab.
29. reference test*.ab.
30. index test*.ab.
31. gold standard.ab.
32. or/4-31.
33. 3 and 32.
38. Limit 33 to "2000 to current".

Web of Science (Web of Knowledge)
#1: TS = (PHQ* OR “Patient Health Questionnaire*”).
#2: TS = (screen* OR prevalence OR “predictive value*” OR detect* OR sensitiv* OR valid* OR revalid* OR predict* OR accura* OR psychometric* OR identif* OR specificit* OR cutoff* OR “cut off*” OR “cut* score*” OR cutpoint* OR “cut point*” OR “threshold score*” OR “reference standard*” OR “reference test*” OR “index test*” OR “gold standard*”).
#1 AND #2.
Indexes = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH.