Psychological Medicine

cambridge.org/psm

Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis[‡]

Yin Wu[1–3], Brooke Levis[1,3], Kira E. Riehm[1], Nazanin Saadat[1], Alexander W. Levis[1], Marleine Azar[1], Danielle B. Rice[1,4], Jill Boruff[5], Pim Cuijpers[6], Simon Gilbody[7], John P.A. Ioannidis[8], Lorie A. Kloda[9], Dean McMillan[7], Scott B. Patten[10,11], Ian Shrier[1,3], Roy C. Ziegelstein[12], Dickens H. Akena[13], Bruce Arroll[14], Liat Ayalon[15], Hamid R. Baradaran[16,17], Murray Baron[1,18], Charles H. Bombardier[19], Peter Butterworth[20,21], Gregory Carter[22], Marcos H. Chagas[23], Juliana C. N. Chan[24–26], Rushina Cholera[27], Yeates Conwell[28], Janneke M. de Man-van Ginkel[29], Jesse R. Fann[30], Felix H. Fischer[31], Daniel Fung[32–35], Bizu Gelaye[36], Felicity Goodyear-Smith[14], Catherine G. Greeno[37], Brian J. Hall[38,39], Patricia A. Harrison[40], Martin Härter[41], Ulrich Hegerl[42], Leanne Hides[43], Stevan E. Hobfoll[44], Marie Hudson[1,18], Thomas Hyphantis[45], Masatoshi Inagaki[46], Nathalie Jetté[10,11,47], Mohammad E. Khamseh[16], Kim M. Kiely[48,49], Yunxin Kwan[50], Femke Lamers[51], Shen-Ing Liu[35,52–54], Manote Lotrakul[55], Sonia R. Loureiro[23], Bernd Löwe[56], Anthony McGuire[57], Sherina Mohd-Sidik[58], Tiago N. Munhoz[59], Kumiko Muramatsu[60], Flávia L. Osório[23,61], Vikram Patel[62,63], Brian W. Pence[64], Philippe Persoons[65,66], Angelo Picardi[67], Katrin Reuter[68], Alasdair G. Rooney[69], Iná S. Santos[59], Juwita Shaaban[70], Abbey Sidebottom[71], Adam Simning[28], Lesley Stafford[72,73], Sharon Sung[32,35], Pei Lin Lynnette Tan[50], Alyna Turner[74,75], Henk C. van Weert[76], Jennifer White[77], Mary A. Whooley[78–80], Kirsty Winkley[81], Mitsuhiko Yamada[82], Andrea Benedetti[3,18,83] and Brett D. Thombs[1–4,18,84]

Abstract

Background. Item 9 of the Patient Health Questionnaire-9 (PHQ-9) queries about thoughts of death and self-harm, but not suicidality. Although it is sometimes used to assess suicide risk, most positive responses are not associated with suicidality. The PHQ-8, which omits Item 9, is thus increasingly used in research. We assessed equivalency of total score correlations and the diagnostic accuracy to detect major depression of the PHQ-8 and PHQ-9.

Methods. We conducted an individual patient data meta-analysis. We fit bivariate random-effects models to assess diagnostic accuracy.

Results. 16 742 participants (2097 major depression cases) from 54 studies were included. The correlation between PHQ-8 and PHQ-9 scores was 0.996 (95% confidence interval 0.996 to 0.996). The standard cutoff score of 10 for the PHQ-9 maximized sensitivity + specificity for the PHQ-8 among studies that used a semi-structured diagnostic interview reference standard ($N = 27$). At cutoff 10, the PHQ-8 was less sensitive by 0.02 (−0.06 to 0.00) and more specific by 0.01 (0.00 to 0.01) among those studies ($N = 27$), with similar results for studies that used other types of interviews ($N = 27$). For all 54 primary studies combined, across all cutoffs, the PHQ-8 was less sensitive than the PHQ-9 by 0.00 to 0.05 (0.03 at cutoff 10), and specificity was within 0.01 for all cutoffs (0.00 to 0.01).

Conclusions. PHQ-8 and PHQ-9 total scores were similar. Sensitivity may be minimally reduced with the PHQ-8, but specificity is similar.

Introduction

The 9-item Patient Health Questionnaire (PHQ-9) (Kroenke *et al.*, 2001) is a self-report questionnaire that is commonly used for identifying people who may have depression based on matching symptoms to diagnostic criteria or, more commonly, on a standard cutoff of a score of 10 or greater (Moriarty *et al.*, 2015; Levis *et al.*, 2019). It is also used as a continuous

Original Article

[‡]The original version of this article was submitted with incorrect author information. A notice detailing this has been published and the error rectified in the online PDF and HTML copies.

Cite this article: Wu Y *et al* (2020). Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychological Medicine* **50**, 1368–1380. https://doi.org/10.1017/S0033291719001314

Received: 25 March 2019
Revised: 10 May 2019
Accepted: 14 May 2019
First published online: 12 July 2019

Key words:
Depression; diagnostic accuracy; individual participant data meta-analysis; meta-analysis; PHQ-8; PHQ-9; screening; systematic review

Authors for correspondence:
Andrea Benedetti, E-mail: andrea.benedetti@mcgill.ca and Brett D. Thombs, E-mail: brett.thombs@mcgill.ca

CAMBRIDGE UNIVERSITY PRESS

measure to assess depressive symptom severity in research and clinical care (Kroenke *et al.*, 2001). The nine items of the PHQ-9 are designed to capture the nine Diagnostic and Statistical Manual of Mental Disorders (DSM) symptom criteria for a major depressive episode (American Psychiatric Association 2013). Response options on the items range from 'not at all' (0 points) to 'nearly every day' (3 points). Per the DSM-5, the ninth criterion for major depression requires 'Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide' (American Psychiatric Association, 2013). Item 9 of the PHQ-9 taps into this criterion but also assesses self-harm, which is not part of the DSM criterion, or passive thoughts of death within the last two weeks: '… how often have you been bothered by…thoughts that you would be better off dead or of hurting yourself in some way?' It does not query specifically about suicidality, and positive responses may be due to thoughts about death or to thoughts about self-harm.

Item 9 is sometimes used as an indicator of suicide risk, and it may be useful as a component of modelling approaches for stratifying suicide risk among participants in psychiatric settings (Simon *et al.*, 2013; 2016). However, responses to the item may not accurately reflect whether or not suicide risk is present, particularly among patients with serious medical conditions for whom thoughts of death may not reflect suicidal ideation, and it appears to perform poorly in identifying individuals at risk in these settings. Four studies in non-psychiatric settings have compared positive responses on Item 9 to responses to questions that explicitly assess suicidal thoughts or intentionality. In these studies, which included US military veterans in primary care (Corson *et al.*, 2004), patients with coronary artery disease (Razykov *et al.*, 2012; Suarez *et al.*, 2015), and cancer patients (Walker *et al.*, 2011), 7% to 21% of all study participants had positive responses on Item 9, but of those, only 18% to 35% had thoughts of suicide based on questions designed specifically to address suicide risk, and only 3% to 20% had a plan (Corson *et al.*, 2004; Walker *et al.*, 2011; Razykov *et al.*, 2012; Suarez *et al.*, 2015). Thus, concerns have been raised that using Item 9 to identify individuals at risk would result in a high rate of false indications, compared to questions designed specifically to assess suicidal thoughts or intentionality (Walker *et al.*, 2011; Razykov *et al.*, 2012; Suarez *et al.*, 2015).

The PHQ-8 omits Item 9 from the PHQ-9. Many research studies use the PHQ-8 as a depression screening tool or to assess depressive symptom severity in order to avoid the high risk of inaccurate indications of suicide risk based on Item 9 (Corson *et al.*, 2004; Kroenke *et al.*, 2009; Razykov *et al.*, 2012; Wells *et al.*, 2013; Barrera *et al.*, 2017). This is a particularly important consideration in studies that are not focused on depression or psychiatric disorders, but would need to allocate substantial resources to follow-up on responses to Item 9 of the PHQ-9. Similarly, many large epidemiological studies that include assessments of depressive symptoms are not able to provide adequate assessment and intervention with telephone or internet surveys (Kroenke *et al.*, 2009).

Although differences in performance between the PHQ-8 and PHQ-9 might be expected to be minimal, to the best of our knowledge, only one study has attempted to verify this by comparing diagnostic accuracy between the PHQ-8 and PHQ-9 (Razykov *et al.*, 2012). That study evaluated the diagnostic testing accuracy of the PHQ-8 *v.* the PHQ-9 and the correlation between PHQ-8 and PHQ-9 scores in a sample of 1022 coronary artery disease outpatients (233 major depression cases). Differences between sensitivity and specificity for the PHQ-8 (50%, 91%) and PHQ-9 (54%, 90%) based on a cutoff score of 10 or greater were minimal. In addition, PHQ-8 and PHQ-9 scores were highly correlated (r = 0.997) (Razykov *et al.*, 2012). One additional study reported correlations between continuous PHQ-8 and PHQ-9 scores (Corson *et al.*, 2004). That study, which included over 1000 patients from a US Department of Veterans Affairs primary care setting, reported a correlation of r = 0.998 (Corson *et al.*, 2004).

We have synthesized a large database of individual participant data (IPD) from primary studies on the PHQ-9 (Levis *et al.*, 2018; 2019). In the present study we included studies from that database that provided individual item scores (not just total PHQ-9 scores), which allowed for calculation of PHQ-8 scores. The objectives of the present study were (1) to evaluate the equivalency of the correlation between PHQ-8 and PHQ-9 scores for assessing depressive symptom severity; and (2) to assess the equivalency of the diagnostic accuracy of PHQ-8 and PHQ-9 across relevant cutoffs for screening to detect major depression.

Method

Data source

The present study used a subset of participants from an IPD database of PHQ-9 (Levis *et al.*, 2018; 2019). The main PHQ-9 IPD meta-analysis (IPDMA) was registered in PROSPERO (CRD42014010673), and a protocol was published (Thombs *et al.*, 2014). Analyses of the diagnostic accuracy of the PHQ-8 were conducted according to protocol with two exceptions: (1) we stratified results by reference standard categories and (2) we added an examination of equivalency with the PHQ-9. Results from the main IPDMA of the PHQ-9 are available elsewhere (Levis *et al.*, 2019).

Search strategy

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations, PsycINFO, and Web of Science from 1 January 2000 through 7 February 2015, using a peer-reviewed search strategy (Canadian Agency for Drugs and Technologies in Health, 2016) (online Supplementary Methods1). We limited our search to these databases based on research showing that adding other databases when the Medline search is highly sensitive does not identify additional eligible studies (Rice *et al.*, 2016). The search was limited to the year 2000 forward because the PHQ-9 was originally published in 2001 (Kroenke *et al.*, 2001). In addition to the database search, we reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada), which was used to store and track search results, conduct screening for eligibility, document correspondence with primary study authors, and extract study characteristics.

Identification of eligible studies

Datasets from articles in any language were eligible for inclusion if they included diagnostic classification among participants aged 18 or older for current Major Depressive Disorder (MDD) or Major Depressive Episode (MDE) based on a validated semi-structured or fully structured interview conducted within two weeks of

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press

PHQ-9 administration, since diagnostic criteria for major depression are for symptoms in the last two weeks. Datasets where not all participants were at least 18 years of age were included if the primary data allowed us to select participants who were at least 18 years of age. Datasets where not all participants were administered the PHQ-9 within two weeks of the diagnostic interview were included if the primary data allowed us to select participants who were administered both instruments within two weeks. Data from studies where the PHQ-9 was administered exclusively to individuals with known psychiatric diagnoses or symptoms or who were seeking psychiatric care were excluded, because screening is not indicated for patients already seeking care or managed in psychiatric settings. For defining major depression cases, we considered MDD or MDE based on the DSM or the International Classification of Diseases (ICD). If more than one was reported, we prioritized MDE over MDD and DSM over ICD. Across all studies, there were only 23 discordant diagnoses that depended on classification prioritization (0.1% of participants). For the present study, we only included primary studies that provided individual PHQ-9 item scores and not just PHQ-9 total scores, because only those datasets allowed us to generate PHQ-8 scores and compare the PHQ-8 with the PHQ-9.

Two investigators independently reviewed titles and abstracts for eligibility. If either reviewer deemed a study potentially eligible, full-text article review was done by two investigators, independently. Disagreement between reviewers after full-text review was resolved by consensus, consulting a third investigator when necessary. Translators were consulted to evaluate titles, abstracts and full-text articles for languages other than those for which team members were fluent.

Data contribution and synthesis

Authors of eligible datasets were invited to contribute de-identified primary data. Primary study country, clinical setting, language, and diagnostic interview administered were extracted from published reports by two investigators independently, with disagreements resolved by consensus. Countries were categorized as 'very high', 'high', or 'low-medium' development level based on the United Nation's human development index (Whiting *et al.*, 2011). Recruitment settings were categorized as 'non-medical', 'primary care', 'inpatient specialty care', or 'outpatient specialty care.' Participant-level data included age, sex, major depression status, current diagnosis or treatment for a mental health problem, and PHQ-9 item scores. In two primary studies, multiple recruitment settings were included, thus recruitment setting was coded at the participant-level. When primary study datasets included appropriate statistical weighting to reflect sampling procedures, we used the provided weights. For studies where sampling procedures merited weighting, but the original study did not weight, we constructed appropriate weights using inverse selection probabilities. Weighting occurred, for instance, when all participants with positive screens, but only a random subset of participants with negative screens, were administered a diagnostic interview.

Individual participant data were converted to a standard format and entered into a single dataset that also included study-level data. We compared published participant characteristics and diagnostic accuracy results with those obtained using the raw datasets. When primary data and original publications were discrepant, we identified and corrected errors when possible and resolved any outstanding discrepancies in consultation with the original investigators.

Statistical analyses

To evaluate the equivalence of the PHQ-8 and PHQ-9 scores for assessing depressive symptom severity, a Pearson correlation with a 95% confidence interval (CI) was calculated between the total scores of PHQ-8 (which excluded Item 9) and PHQ-9.

To estimate the diagnostic accuracy of the PHQ-8 and compare with the PHQ-9, we analyzed primary studies separately by the type of diagnostic interview that was used as the reference standard, as we did in the previously published main PHQ-9 meta-analysis (Levis *et al.*, 2019). This was done because of differences in the performance of the different types of interviews. Semi-structured interviews involve clinical judgement and are designed to be administered by clinically trained professionals; fully structured interviews are completely scripted and designed for lay administration, but the resulting increased standardization and reliability across interviewers may lead to increased misclassification (Brugha *et al.*, 1999; Nosen and Woody, 2008). The Mini International Neuropsychiatric Interview (MINI), which is a fully structured interview, was developed to be administered in a fraction of the time necessary for other fully structured interviews and was described by its developers as designed to be over-inclusive (Robins *et al.*, 1988; Sheehan *et al.*, 1997). In a previous study, we found that semi-structured and fully structured diagnostic interviews are not interchangeable reference standards for major depression and that fully structured interviews may diagnose depression at higher rates than semi-structured interviews at low symptom levels and at lower rates at high symptom levels (Levis *et al.*, 2018). We also found that the MINI classifies approximately twice as many participants as cases compared to the most commonly used fully-structured interview, the Composite International Diagnostic Interview (CIDI) (Levis *et al.*, 2018). In the main PHQ-9 meta-analysis, the diagnostic accuracy of the PHQ-9 differed substantially depending on the reference standard used for the comparison (Levis *et al.*, 2019). Thus, for the present study, we analyzed primary studies separately based on whether they used a semi-structured interview, a fully structured interview (non-MINI), or the MINI.

For each reference standard and for the PHQ-8 and PHQ-9 cutoffs 5–15, separately, bivariate random-effects models were fitted using an adaptive Gauss Hermite quadrature with 1 quadrature point (Riley *et al.*, 2008). This 2-stage meta-analytic approach models sensitivity and specificity at the same time, taking the inherent correlation between them and the precision of estimates within studies into account. A random-effects model was used as we assumed true values of sensitivity and specificity would likely to vary across primary studies.

In order to examine the equivalence between PHQ-8 and PHQ-9 across reference standards, for each analysis, we used the results of the random-effects meta-analyses at each cutoff to construct separate empirical receiver operating characteristic (ROC) curves based on the pooled estimates. Equivalence tests between PHQ-8 and PHQ-9 sensitivity and specificity were conducted at each cutoff. This allowed us to test whether the sensitivity and specificity of the PHQ-8 was similar to that of the PHQ-9, up to a prespecified maximum clinically acceptable difference, that is, an equivalence margin (Walker and Nowacki, 2011). In the present study, an equivalence margin of $\delta = 0.05$ was used, which is the same margin that was used in a previous study that used the same IPD database (Ishihara *et al.*, 2019). CIs for the differences between PHQ-8 and PHQ-9 sensitivity and specificity at each cutoff were constructed via a cluster bootstrap approach (van der Leeden *et al.*, 1997; 2008), with resampling at the study and subject

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press

level. For each comparison, we ran 1000 iterations of the bootstrap. For each bootstrap iteration, the bivariate random-effects model was fitted to the PHQ-8 and PHQ-9 data, and the pooled sensitivities and specificities were computed separately, as described above, for each cutoff score. Equivalence tests were done by comparing the CIs around the pooled sensitivity and specificity differences to the equivalence margin of $\delta = 0.05$. If the entire CI was included within the interval of +/− 0.05, then we rejected the hypothesis that there were differences large enough to be important and concluded that equivalence was present. If the entire CI was outside of the interval, then we failed to reject the hypothesis that the PHQ-8 and PHQ-9 were not equivalent. When the CIs crossed the +/− 0.05 threshold, findings were deemed equivocal, and the equivalence was indeterminate.

Although we previously found that sensitivity and specificity of the PHQ-9 differs by type of reference standard, we did not believe that the differences in sensitivity and specificity between the PHQ-8 and PHQ-9 would vary depending on the reference standard. This is because for each included study the PHQ-8 and PHQ-9 were compared to the same reference standard. Thus, we reported pooled sensitivity and specificity only stratified by reference standards, but we investigated equivalence both stratified by reference standards and pooled across all studies. To investigate heterogeneity across studies, by reference standard and overall, we generated forest plots for the differences in sensitivity and specificity estimates between PHQ-8 and PHQ-9 for the standard cutoff 10 for each study. We also quantified heterogeneity at cutoff 10, by reporting the estimated variances of the random effects for the differences in PHQ-8 and PHQ-9 sensitivity and specificity (τ^2) (Higgins and Thompson, 2002; Fagerland *et al.*, 2014).

All analyses were run in R (R version R 3.5.0 and R Studio version 1.1.423) using the lme4 package.

Results

Search results and inclusion of primary data

For the main IPDMA, of 5248 unique titles and abstracts identified from the database search, 5039 were excluded after title and abstract review and 113 after full-text (online Supplementary List1), leaving 96 eligible articles with data from 69 unique participant samples (online Supplementary Fig.1). Of the 69 unique samples, 55 contributed data (80%). In addition, authors of included studies contributed data from three unpublished studies, for a total of 58 PHQ-9 datasets contributed to our IPDMA. Four studies without PHQ-9 individual item scores were excluded from the present study (see online Supplementary Table1b). Thus, 16 742 participants (2097 major depression cases) from 54 studies were analyzed (78% of 21 572 participants from the 69 eligible published studies and 3 eligible unpublished studies). Included study characteristics are shown in online Supplementary Table1a. Characteristics of eligible studies that did not provide data, including the 4 studies excluded because they only provided PHQ-9 total scores, are shown in online Supplementary Table1b.

There were 27 included primary studies that used semi-structured interviews to assess major depression (6362 participants), 13 that used fully structured interviews other than the MINI (7596 participants), and 14 that used the MINI (2784 participants). The Structured Clinical Interview for DSM Disorders (SCID) was the most commonly used semi-structured interview (24 studies, 4378 participants), and the CIDI was the most commonly used fully structured interview (10 studies, 6291 participants). The average study sample size and number of major depression cases was 236 and 29 for studies that used a semi-structured interview; 584 and 61 for studies that used a fully structured interview; and 199 and 37 for studies that used the MINI.

Participant characteristics are shown in Table 1.

Table 1. Participant characteristics by subgroup

Participant subgroup	*N* Participants	*N* (%) Major depression
All participants	16 742	2097 (13)
Type of diagnostic interview		
Semi-structured diagnostic interview	6362	790 (12)
Fully structured diagnostic interview	7596	790 (10)
Mini International Neuropsychiatric Interview	2784	517 (19)
Age[a]		
<60	11 144	1402 (13)
⩾60	5552	692 (12)
Sex[a]		
Women	9552	1259 (13)
Men	7180	835 (12)
Care setting		
Non-medical care	1832	252 (14)
Primary care	7846	760 (10)
Inpatient specialty care	1245	136 (11)
Outpatient specialty care	5819	949 (16)
Country human development index		
Very high	13 297	1577 (12)
High	1337	276 (21)
Low-medium	2108	244 (12)

[a]Due to missing participant data, total participant numbers for these variables were <16 742.

Table 2. Characteristics of participants who rated Item 9 as present for several days, more than half the days, or nearly every day (i.e. scores 1–3) in last two weeks by total PHQ-8 score

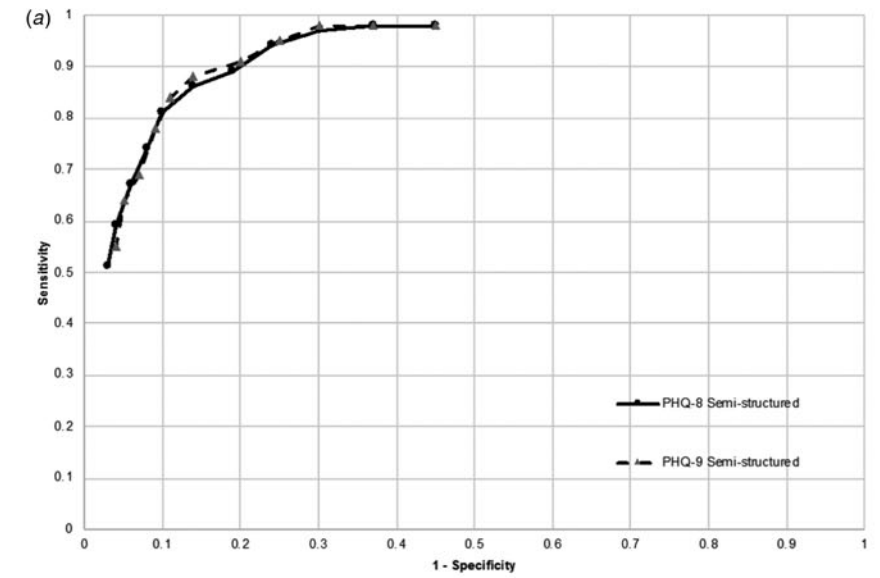
Total PHQ-8 score	*N* of participants[a]	% with non-zero Item 9	Item 9 mean (s.d.)
0–4	11 034	1.9	0.02 (0.16)
5–9	5071	13.2	0.16 (0.45)
10–14	2231	31.3	0.44 (0.74)
15–19	1044	48.3	0.78 (0.97)
20–24	380	64.7	1.39 (1.25)

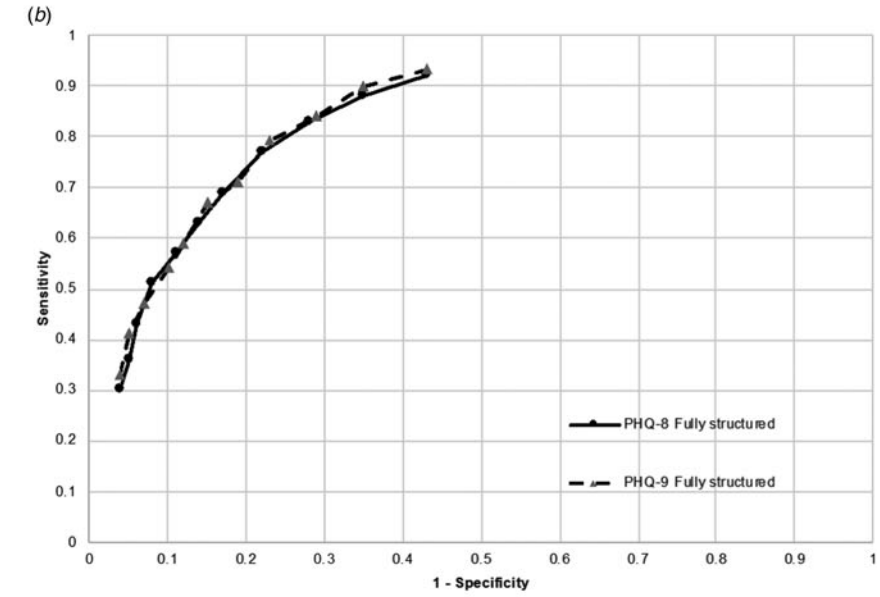
[a]Numbers of participants add up to >16 742 as they were weighted by sampling weights.

PHQ-8 and PHQ-9 scores

Among all participants in all studies, the mean [standard deviation (s.d.)] PHQ-8 score (range = 0–24) was 5.3 (5.2),

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press





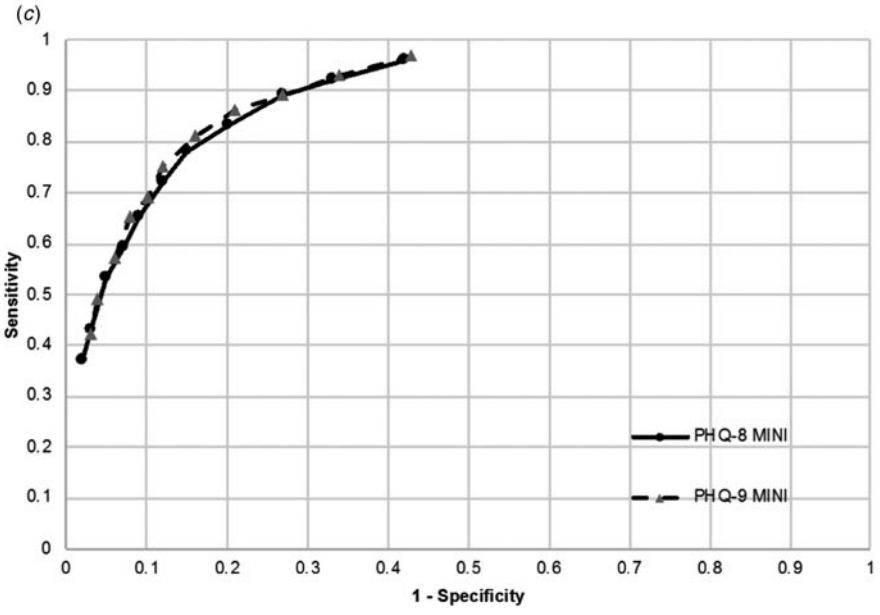


Fig. 1. (*a*) ROC curves for PHQ-8 and PHQ-9 among studies that used a semi-structured reference standard. (*b*) ROC curves for PHQ-8 and PHQ-9 among studies that used a fully structured reference standard (MINI excluded). (*c*) ROC curves for PHQ-8 and PHQ-9 among studies that used the MINI reference standard.

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press

Table 3. Comparison of sensitivity and specificity estimates between PHQ-8 and PHQ-9 among studies that used a semi-structured reference standard

	PHQ-8[a]				PHQ-9				PHQ-8 – PHQ-9[b]			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.98	(0.95–0.99)	0.55	(0.50–0.60)	0.98	(0.95–0.99)	0.55	(0.50–0.60)	0.00	(−0.01 to 0.00)	0.00	(0.00–0.01)
6	0.98	(0.95–0.99)	0.63	(0.58–0.68)	0.98	(0.95–0.99)	0.63	(0.58–0.67)	0.00	(−0.00 to 0.00)	0.00	(0.00–0.01)
7	0.97	(0.93–0.99)	0.70	(0.66–0.74)	0.98	(0.93–0.99)	0.70	(0.65–0.74)	−0.01	(−0.02 to 0.00)	0.00	(0.00–0.01)
8	0.94	(0.89–0.96)	0.76	(0.72–0.79)	0.95	(0.90–0.97)	0.75	(0.71–0.79)	−0.01	(−0.03 to 0.00)	0.01	(0.00–0.01)
9	0.89	(0.84–0.92)	0.81	(0.78–0.84)	0.91	(0.87–0.95)	0.80	(0.77–0.83)	−0.02	(−0.06 to −0.00)	0.01	(0.00–0.01)
10[b]	0.86	(0.80–0.90)	0.86	(0.83–0.89)	0.88	(0.82–0.92)	0.86	(0.82–0.88)	−0.02	(−0.06 to −0.00)	0.00	(0.00–0.02)
11	0.81	(0.75–0.87)	0.90	(0.87–0.92)	0.84	(0.84–0.84)	0.89	(0.89–0.89)	−0.03	(−0.06 to −0.00)	0.01	(0.00–0.02)
12	0.74	(0.68–0.79)	0.92	(0.89–0.93)	0.78	(0.71–0.83)	0.91	(0.89–0.93)	−0.04	(−0.09 to −0.01)	0.01	(0.00–0.01)
13	0.67	(0.60–0.73)	0.94	(0.92–0.95)	0.69	(0.63–0.75)	0.93	(0.91–0.95)	−0.02	(−0.07 to −0.00)	0.01	(0.00–0.01)
14	0.59	(0.53–0.65)	0.96	(0.94–0.97)	0.64	(0.57–0.70)	0.95	(0.93–0.96)	−0.05	(−0.09 to −0.01)	0.01	(0.00–0.01)
15	0.51	(0.44–0.57)	0.97	(0.95–0.98)	0.55	(0.48–0.62)	0.96	(0.94–0.97)	−0.04	(−0.09 to −0.02)	0.01	(0.00–0.01)

CI, confidence interval.
[a]*N* Studies = 27; *N* Participants = 6362; *N* major depression = 790.
[b]For PHQ-8 cutoff 10, among studies that used semi-structured interviews as the reference standard, the default optimizer in glmer failed to converge, thus bobyqa was used instead.

Table 4. Comparison of sensitivity and specificity estimates between PHQ-8 and PHQ-9 among studies that used a fully structured reference standard (MINI excluded)

	PHQ-8[a]				PHQ-9				PHQ-8 – PHQ-9			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.92	(0.85–0.96)	0.57	(0.49–0.66)	0.93	(0.85–0.96)	0.57	(0.48–0.65)	−0.01	(−0.03 to 0.00)	0.00	(0.00–0.02)
6	0.88	(0.79–0.93)	0.65	(0.57–0.73)	0.90	(0.81–0.95)	0.65	(0.56–0.72)	−0.02	(−0.07 to 0.00)	0.00	(0.00–0.02)
7	0.83	(0.73–0.90)	0.72	(0.64–0.79)	0.84	(0.73–0.90)	0.71	(0.63–0.78)	−0.01	(−0.01 to 0.00)	0.01	(0.00–0.02)
8	0.77	(0.66–0.85)	0.78	(0.71–0.84)	0.79	(0.68–0.86)	0.77	(0.70–0.83)	−0.02	(−0.07 to −0.00)	0.01	(0.00–0.01)
9	0.69	(0.59–0.77)	0.83	(0.76–0.87)	0.71	(0.62–0.80)	0.81	(0.75–0.86)	−0.02	(−0.07 to −0.00)	0.02	(0.01–0.02)
10	0.63	(0.52–0.72)	0.86	(0.81–0.90)	0.67	(0.57–0.76)	0.85	(0.80–0.90)	−0.04	(−0.09 to −0.01)	0.01	(0.00–0.02)
11	0.57	(0.45–0.67)	0.89	(0.85–0.93)	0.59	(0.49–0.69)	0.88	(0.84–0.92)	−0.02	(−0.07 to −0.01)	0.01	(0.00–0.02)
12	0.51	(0.38–0.64)	0.92	(0.88–0.94)	0.54	(0.43–0.65)	0.90	(0.86–0.93)	−0.03	(−0.16 to −0.01)	0.02	(0.01–0.02)
13	0.43	(0.32–0.55)	0.94	(0.91–0.96)	0.47	(0.36–0.58)	0.93	(0.89–0.95)	−0.04	(−0.12 to −0.01)	0.01	(0.00–0.01)
14	0.36	(0.26–0.47)	0.95	(0.93–0.97)	0.41	(0.31–0.53)	0.95	(0.92–0.96)	−0.05	(−0.14 to −0.01)	0.00	(0.00–0.01)
15	0.30	(0.22–0.39)	0.96	(0.95–0.98)	0.33	(0.24–0.42)	0.96	(0.94–0.97)	−0.03	(−0.07 to −0.00)	0.00	(0.00–0.00)

CI, confidence interval.
[a]*N* Studies = 13; *N* Participants = 7596; *N* major depression = 790.

Table 5. Comparison of sensitivity and specificity estimates between PHQ-8 and PHQ-9 among studies that used the MINI reference standard

	PHQ-8[a]				PHQ-9				PHQ-8 – PHQ-9			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.96	(0.93–0.98)	0.58	(0.50–0.65)	0.97	(0.93–0.98)	0.57	(0.49–0.65)	−0.01	(−0.01 to 0.00)	0.01	(0.00–0.02)
6	0.92	(0.85–0.96)	0.67	(0.59–0.74)	0.93	(0.86–0.97)	0.66	(0.59, 0.73)	−0.01	(−0.03 to 0.00)	0.01	(0.00–0.02)
7	0.89	(0.81–0.94)	0.73	(0.67–0.79)	0.89	(0.81–0.94)	0.73	(0.66–0.79)	0.00	(−0.03 to 0.00)	0.01	(0.00–0.01)
8	0.83	(0.75–0.89)	0.80	(0.75–0.84)	0.86	(0.77–0.91)	0.79	(0.74–0.83)	−0.03	(−0.06 to 0.00)	0.01	(0.00–0.02)
9	0.78	(0.69–0.85)	0.85	(0.81–0.89)	0.81	(0.71–0.88)	0.84	(0.80–0.88)	−0.03	(−0.07 to −0.00)	0.01	(0.00–0.02)
10	0.72	(0.63–0.79)	0.88	(0.84–0.91)	0.75	(0.66–0.82)	0.88	(0.84–0.91)	−0.03	(−0.08 to −0.01)	0.01	(0.00–0.02)
11	0.65	(0.57–0.73)	0.91	(0.88–0.94)	0.69	(0.61–0.77)	0.90	(0.87–0.93)	−0.04	(−0.08 to −0.01)	0.01	(0.00–0.02)
12	0.59	(0.51–0.66)	0.93	(0.91–0.95)	0.65	(0.56–0.73)	0.92	(0.90–0.94)	−0.06	(−0.11 to −0.02)	0.01	(0.00–0.02)
13	0.53	(0.44–0.62)	0.95	(0.93–0.97)	0.57	(0.49–0.66)	0.94	(0.92–0.96)	−0.04	(−0.09 to −0.01)	0.01	(0.00–0.02)
14	0.43	(0.35–0.51)	0.97	(0.95–0.98)	0.49	(0.49–0.49)	0.96	(0.96–0.96)	−0.06	(−0.11 to −0.02)	0.01	(0.00–0.02)
15	0.37	(0.29–0.45)	0.98	(0.96–0.99)	0.42	(0.42–0.42)	0.97	(0.97–0.97)	−0.05	(−0.10 to −0.02)	0.01	(0.00–0.01)

CI, confidence interval; MINI, Mini International Neuropsychiatric Interview.
[a]*N* Studies = 14; *N* Participants = 2784; *N* major depression = 517.

Table 6. Comparison of sensitivity and specificity estimates between PHQ-8 and PHQ-9 across cutoffs 5–15 for all studies

	PHQ-8 – PHQ-9			
Cutoff	Sensitivity	95% CI	Specificity	95% CI
5	−0.01	(−0.01 to 0.00)	0.00	(0.00–0.01)
6	0.00	(−0.01 to 0.00)	0.01	(0.00–0.01)
7	−0.01	(−0.02 to 0.00)	0.00	(0.00–0.01)
8	−0.01	(−0.04 to −0.01)	0.00	(0.01–0.01)
9	−0.03	(−0.06 to −0.01)	0.01	(0.01–0.01)
10	−0.03	(−0.06 to −0.02)	0.01	(0.00–0.01)
11	−0.03	(−0.06 to −0.01)	0.01	(0.01–0.01)
12	−0.05	(−0.08 to −0.03)	0.01	(0.00–0.01)
13	−0.04	(−0.06 to −0.02)	0.00	(0.00–0.01)
14	−0.05	(−0.08 to −0.03)	0.01	(0.00–0.01)
15	−0.04	(−0.07 to −0.03)	0.01	(0.00–0.01)

CI, confidence interval.
[a]*N* Studies = 54; *N* Participants = 16 742; *N* major depression = 2097.

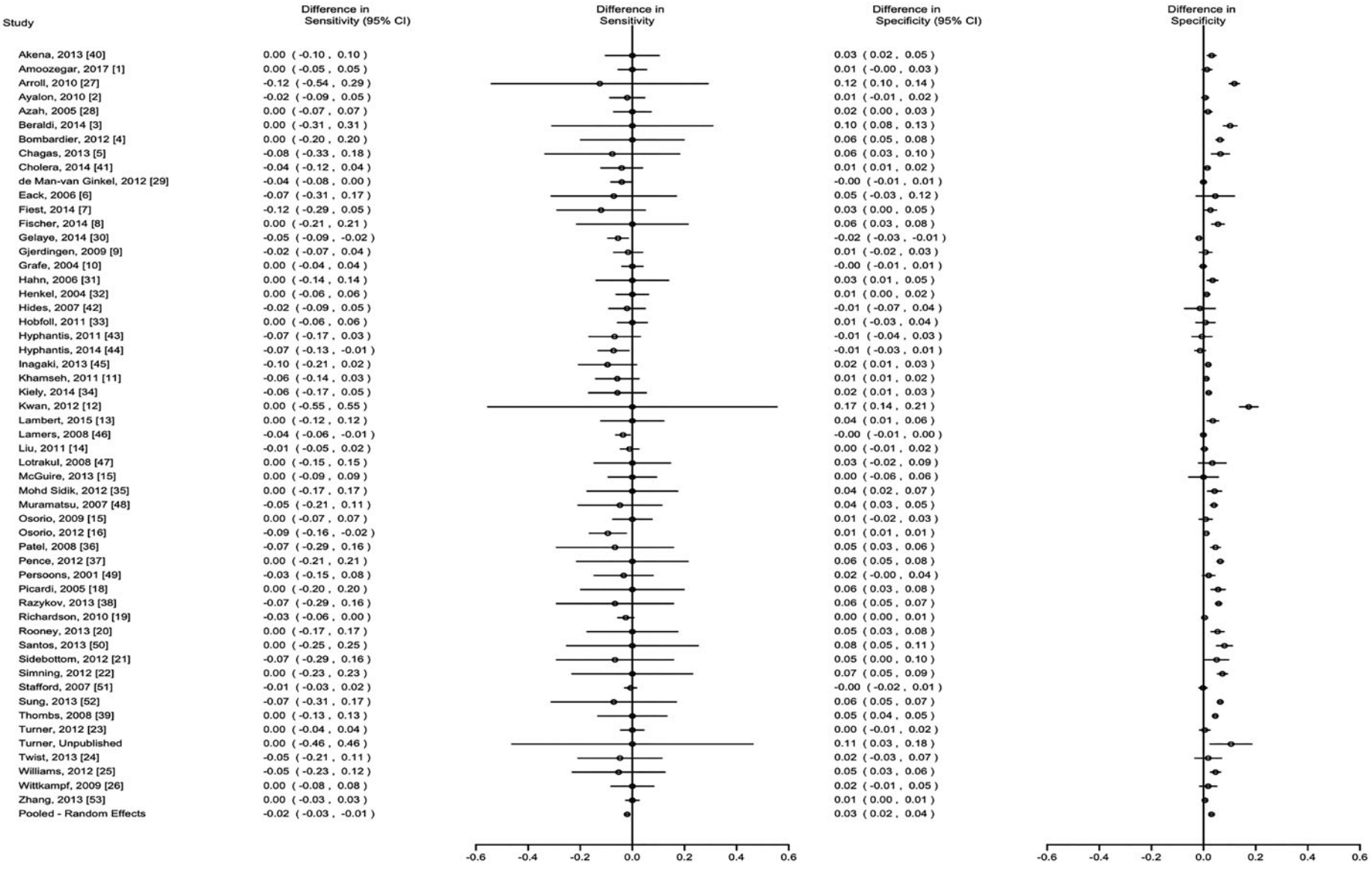


Fig. 2. Forest plots of the difference in sensitivity and specificity estimates at cutoff 10 between PHQ-8 and PHQ-9 among all studies[a] (*N* Studies = 54[b]; *N* Participants = 16 742; *N* major depression = 2097)[c]. [a] τ^2 for the difference of sensitivity and specificity were both <0.001. [b] The reference numbers refer to online Supplementary Material References. [c] Amoozegar, 2017 [1] and Lambert, 2015 [13] were unpublished at the time of the electronic database search then subsequently published.

and the mean (s.d.) PHQ-9 score (range 0–27) was 5.4 (5.4). Overall, 11.8% of participants had a non-zero score on Item 9 (score of 1–3). As shown in Table 2, this included 1.9% among participants with PHQ-8 scores 0–4 and increased to 64.7% among those with scores 20–24. The correlation (95% CI) between PHQ-8 and PHQ-9 scores was 0.996 (0.996, 0.996). The correlation of the score of Item 9 with PHQ-8 scores was 0.480 (0.469, 0.492).

Diagnostic accuracy of the PHQ-8 and PHQ-9

ROC curves comparing sensitivity and specificity estimates for cutoffs 5–15 between the PHQ-8 and PHQ-9 for the three reference standard categories, separately, are shown in Fig. 1. The curves for the PHQ-8 and PHQ-9 were highly overlapping for each reference standard, and the area under the curve for the PHQ-8 and PHQ-9 were similar for semi-structured interviews (0.930 *v.* 0.933), fully structured interviews (excluding the MINI; 0.852 *v.* 0.856), and the MINI (0.894 *v.* 0.899).

Comparisons of sensitivity and specificity estimates between PHQ-8 and PHQ-9 cutoffs 5–15 across the three reference standard categories are shown in Tables 3–5. Cutoff 10 maximized combined sensitivity and specificity for PHQ-8 [sensitivity (95% CI) = 0.86 (0.80–0.90), specificity (95% CI) = 0.86 (0.83–0.89)] and PHQ-9 [sensitivity (95% CI) = 0.88 (0.82–0.92), specificity (95% CI) = 0.86 (0.82–0.88)] among studies using a semi-structured interview as the reference standard; cutoff 8 for PHQ-8 [sensitivity (95% CI) = 0.77 (0.66–0.85), specificity (95% CI) = 0.78 (0.71–0.84)] and PHQ-9 [sensitivity (95% CI) = 0.79 (0.68–0.86), specificity (95% CI) = 0.77 (0.70–0.83)] among studies using a fully structured interview; and cutoff 8 for PHQ-8 [sensitivity (95% CI) = 0.83 (0.75–0.89), specificity (95% CI) = 0.80 (0.75–0.84)] and PHQ-9 [sensitivity (95% CI) = 0.86 (0.77–0.91), specificity (95% CI) = 0.79 (0.74–0.83)] among studies using the MINI.

In comparisons stratified by reference standard, for sensitivity, results of equivalence tests showed that for semi-structured diagnostic interviews, estimates were equivalent from cutoffs 5 through 9 and indeterminate from cutoffs 10 through 15; for fully structured interviews (excluding the MINI), they were equivalent on cutoffs 5 and 7 and indeterminate at cutoffs 6 and 8 through 15; and for the MINI, they were equivalent from cutoffs 5 through 7 and indeterminate from cutoffs 8 through 15. Estimates of specificity were equivalent in all analyses, regardless of reference standards and cutoffs. See Tables 3–5.

Overall, including all 54 primary studies, as shown in Table 6, across cutoffs, sensitivity was between 0.00 and 0.05 percentage points lower for the PHQ-8 compared to the PHQ-9. At cutoff 10, the difference (95% CI) was −0.03 (−0.06 to −0.02). For specificity, the PHQ-8 and PHQ-9 were within 0.01 for all cutoffs. For sensitivity, estimates were equivalent for cutoffs 5 to 8 and indeterminate for cutoffs 9 to 15. For specificity, estimates were equivalent for all cutoffs.

A forest plot of the difference of sensitivity and specificity estimates for cutoff 10 between PHQ-8 and PHQ-9 for all studies is shown in Fig. 2. At the commonly used cutoff of 10 or greater, there was low heterogeneity in the differences across the 54 studies with estimated inter-study heterogeneity (τ^2) <0.01 for sensitivity and <0.01 for specificity. Forest plots of the differences of sensitivity and specificity estimates for cutoff 10 between PHQ-8 and PHQ-9 among studies by reference standard category are shown in online Supplementary Fig. 2.

Discussion

In the present study, we assessed the correlation of continuous PHQ-8 and PHQ-9 scores for assessing depression severity in research and clinical practice, and we compared the diagnostic accuracy of the PHQ-8 and PHQ-9 across all cutoffs for detecting major depression. There were two main findings. First, the correlation of continuous PHQ-8 and PHQ-9 scores was high (0.996). Second, to screen for major depression, the PHQ-8 at different possible cutoffs including the standard cutoff of 10 or greater, was similarly accurate compared to the PHQ-9 overall and across all three types of reference standards. The cutoffs that maximized combined sensitivity and specificity were the same for the PHQ-8 and PHQ-9 across reference standard categories.

Overall, for all 54 primary studies combined, across all cutoffs, the PHQ-8 was slightly less sensitive than the PHQ-9 by 0.00 to 0.05 (0.03 at cutoff 10). For specificity, the differences between PHQ-8 and PHQ-9 were within 0.01 for all cutoffs. Although the CIs for the difference in sensitivity for cutoff 10 did not fit the study definition of equivalency, the reduction in sensitivity if the PHQ-8 is used is small, and specificity is equivalent.

Previous studies have shown that Item 9 of the PHQ-9 does not accurately assess suicide risk and identifies far more patients or study participants as at risk than would be identified with items designed to assess suicide risk (Corson *et al.*, 2004; Walker *et al.*, 2011; Razykov *et al.*, 2012). Thus, unintended consequences of using Item 9 could include substantial additional costs for research, as well as possible harms or inconvenience to patients. Research ethics boards sometimes require follow-up for all patients with positive responses to Item 9. Using the PHQ-8, which is minimally different from PHQ-9 in terms of diagnostic accuracy characteristics, would reduce unintended consequences of false signals of suicide risk without substantive changes to continuous measurement properties or diagnostic accuracy for major depression.

It is possible that use of the PHQ-8 could result in not identifying a small subset of people with suicidal thoughts, although, if the case, based on our findings, this number would be small. Furthermore, there is no evidence that using questionnaires to screen for suicide in general medical settings, above and beyond screening for depression, would reduce risk of suicide (Allaby, 2010; Crawford *et al.*, 2011; Siu and the US Preventive Services Task Force, 2016). Tools are available to screen patients or stratify by risk for suicidality. However, a review concluded that they are not accurate enough at this point for use in practice and that alternative methods are more appropriate (Carter and Spittal, 2018). Indeed, in mental health settings or when there is reason to suspect possible suicidality, standards of care indicate that engagement with patients is needed to assess suicide risk and determine the best management plan, as appropriate (Carter and Spittal, 2018).

To our knowledge, this is the first meta-analysis and also the first study using a large IPD database to compare diagnostic accuracy characteristics of PHQ-8 and PHQ-9. Strengths of this study included the large overall sample size, the ability to compare results for PHQ-8 and PHQ-9 from all cutoffs from all studies (rather than just published cutoff results), and the ability to assess diagnostic accuracy separately in studies that used semi- and fully structured diagnostic interviews as the reference standard. There are also limitations to consider. First, for the full IPDMA data, we were unable to include primary data from 14 of 69 published eligible datasets (20% datasets; 17% of eligible participants), and

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press

we restricted our analyses to those with complete data for all individual PHQ-9 item scores (95% of available data). Nonetheless, this sample was much larger than the few previous primary studies that have compared the PHQ-8 and PHQ-9. Second, we categorized studies based on the diagnostic interview that was used, but adaptations to interviews are sometimes made and, thus, all studies may have not used the diagnostic interviews in the way that they were originally designed. However, when we analyzed data from all studies, regardless of reference standard, heterogeneity was minimal, suggesting that findings can be applied across reference standards.

Conclusions

In summary, although the PHQ-9 was designed to reflect the 9 symptoms included in DSM criteria for major depression, the item assessing suicide risk also assesses self-harm. This study used a large IPD dataset and found that the PHQ-8 performs similarly to the PHQ-9 in terms of the correlation of continuous scores and the diagnostic accuracy across all cutoffs for detecting major depression. Removing Item 9 and using the PHQ-8 instead of the PHQ-9 has minimal influence on performance of the measure and will likely reduce the number of false positive signals from people who endorse this item but would not be considered to be at risk for suicide based on measures intended to assess suicide risk.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291719001314

Author ORCIDs. Brett D. Thombs, 0000-0002-5644-8432

Acknowledgements. This study was funded by the Canadian Institutes of Health Research (KRS-134297; PCG-155468). Dr Wu was supported by an Utting Postdoctoral Fellowship from the Jewish General Hospital, Montreal, Quebec. Drs Benedetti and Thombs were supported by Fonds de recherche du Québec – Santé (FRQS) researcher salary awards. Ms. Levis was supported by a CIHR Frederick Banting and Charles Best Canada Graduate Scholarship doctoral award. Ms. Riehm and Ms. Saadat were supported by CIHR Frederick Banting and Charles Best Canadian Graduate Scholarships – Master's Awards. Mr. Levis and Ms. Azar were supported by FRQS Masters Training Awards. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. Collection of data for the study by Arroll *et al.* was supported by a project grant from the Health Research Council of New Zealand. Data collection for the study by Ayalon *et al.* was supported from a grant from Lundbeck International. The primary study by Khamseh *et al.* was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary study by Bombardier *et al.* was supported by the Department of 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. H133N060003), and University of Michigan (grant no. H133N060032). Dr Butterworth was supported by Australian Research Council Future Fellowship FT130101444. Dr Cholera was supported by a United States National Institute of Mental Health (NIMH) grant (5F30MH096664), and the United States National Institutes of Health (NIH) Office of the Director, Fogarty International Center, Office of AIDS Research, National Cancer Center, National Heart, Blood, and Lung Institute, and the NIH Office of Research for Women's Health through the Fogarty Global Health Fellows Program Consortium (1R25TW00934001) and the American Recovery and Reinvestment Act. Dr Conwell received support from NIMH (R24MH071604) and the Centers for Disease Control and Prevention (R49 CE002093). The primary studies by Amoozegar and by Fiest *et al.* were funded by the Alberta Health Services, the University of Calgary Faculty of Medicine, and the Hotchkiss Brain Institute. The primary study by Fischer *et al.* was funded by the German Federal Ministry of Education and Research (01GY1150). Dr Fischler was supported by a grant from the Belgian Ministry of Public Health and Social Affairs and a restricted grant from Pfizer Belgium. Data for the primary study by Gelaye *et al.* was supported by grant from the NIH (T37 MD001449). Collection of data for the primary study by Gjerdingen *et al.* was supported by grants from the NIMH (R34 MH072925, K02 MH65919, P30 DK50456). The primary study by Eack *et al.* was funded by the NIMH (R24 MH56858). Collection of data provided by Drs Härter and Reuter was supported by the Federal Ministry of Education and Research (grants No. 01 GD 9802/4 and 01 GD 0101) and by the Federation of German Pension Insurance Institute. Collection of data for the primary study by Hobfoll *et al.* was made possible in part from grants from NIMH (RO1 MH073687) and the Ohio Board of Regents. Dr Hall received support from a grant awarded by the Research and Development Administration Office, University of Macau (MYRG2015-00109-FSS). The primary study by Hides *et al.* was funded by the Perpetual Trustees, Flora and Frank Leith Charitable Trust, Jack Brockhoff Foundation, Grosvenor Settlement, Sunshine Foundation and Danks Trust. The primary study by Henkel *et al.* was funded by the German Ministry of Research and Education. Data for the study by Razykov *et al.* was collected by the Canadian Scleroderma Research Group, which was funded by the CIHR (FRN 83518), the Scleroderma Society of Canada, the Scleroderma Society of Ontario, the Scleroderma Society of Saskatchewan, Sclérodermie Québec, the Cure Scleroderma Foundation, Inova Diagnostics Inc., Euroimmun, FRQS, the Canadian Arthritis Network, and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montreal, QC. Dr Hudson was supported by a FRQS Senior Investigator Award. Collection of data for the primary study by Hyphantis *et al.* was supported by grant from the National Strategic Reference Framework, European Union, and the Greek Ministry of Education, Lifelong Learning and Religious Affairs (ARISTEIA-ABREVIATE, 1259). The primary study by Inagaki *et al.* was supported by the Ministry of Health, Labour and Welfare, Japan. Dr Jetté was supported by a Canada Research Chair in Neurological Health Services Research. Collection of data for the primary study by Kiely *et al.* was supported by National Health and Medical Research Council (grant number 1002160) and Safe Work Australia. Dr Kiely was supported by funding from an Australian National Health and Medical Research Council fellowship (grant number 1088313). The primary study by Lamers *et al.* was funded by the Netherlands Organisation for Health Research and development (grant number 945-03-047). The primary study by Liu *et al.* was funded by a grant from the National Health Research Institute, Republic of China (NHRI-EX97-9706PI). The primary study by Lotrakul *et al.* was supported by the Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (grant number 49086). 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.* The primary study by Mohd-Sidik *et al.* was funded under the Research University Grant Scheme from Universiti Putra Malaysia, Malaysia and the Postgraduate Research Student Support Accounts of the University of Auckland, New Zealand. The primary study by Santos *et al.* was funded by the National Program for Centers of Excellence (PRONEX/FAPERGS/CNPq, Brazil). The primary study by Muramatsu *et al.* was supported by an educational grant from Pfizer US Pharmaceutical Inc. Collection of primary data for the study by Dr Pence was provided by NIMH (R34MH084673). The primary studies by Osório *et al.* were funded by Reitoria de Pesquisa da Universidade de São Paulo (grant number 09.1.01689.17.7) and Banco Santander (grant number 10.1.01232.17.9). Dr Osório 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. Dr Persoons was supported by a grant from the Belgian Ministry of Public Health and Social Affairs and a restricted grant from Pfizer Belgium. Dr Shaaban was supported by funding from Universiti Sains Malaysia. 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 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 R40MC07840). Simning *et al.*'s research was supported in part by grants from the NIH (T32 GM07356), Agency for Healthcare Research and Quality (R36 HS018246), NIMH (R24 MH071604), and the National Center for

Research Resources (TL1 RR024135). Dr Stafford received PhD scholarship funding from the University of Melbourne. Collection of data for the studies by Turner *et al.* were funded by a bequest from Jennie Thomas through the Hunter Medical Research Institute. 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 Thombs *et al.* was done with data from the Heart and Soul Study (PI Mary Whooley). The Heart and Soul Study was funded by the Department of Veterans Epidemiology Merit Review Program, the Department of Veterans Affairs Health Services Research and Development service, the National Heart Lung and Blood Institute (R01 HL079235), the American Federation for Aging Research, the Robert Wood Johnson Foundation, and the Ischemia Research and Education Foundation. Dr Thombs was supported by an Investigator Award from the Arthritis Society. 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 study by Wittkampf *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. Collection of data for the primary study by Zhang *et al.* was supported by the European Foundation for Study of Diabetes, the Chinese Diabetes Society, Lilly Foundation, Asia Diabetes Foundation and Liao Wun Yuk Diabetes Memorial Fund. No other authors reported funding for primary studies or for their work on the present study.

Author contributions. YW, BLevis, JB, PC, SG, JPAI, LAK, DM, SBP, IS, RCZ, AB, and BDT were responsible for the study conception and design. JB and LAK designed and conducted database searches to identify eligible studies. DHA, BA, LA, HRB, MB, CHB, PB, GC, MHC, JCNC, RC, YC, JMG, JRF, FHF, DF, BG, FGS, CGG, BJH, PAH, MHärter, UH, LH, SEH, MHudson, TH, MI, NJ, MEK, KMK, YK, FL, SL, ML, SRL, BLöwe, AM, SM, TNM, KM, FLO, VP, BWP, PP, AP, KR, AGR, ISS, JS, ASidebottom, ASimning, LS, SS, PLLT, AT, HCvW, JW, MAW, KW, MY, and BDT were responsible for collection of primary data included in this study. BLevis, KER, NS, AWL, MA, DBR, and BDT contributed to data extraction and coding for the meta-analysis. YW, BLevis, AB, and BDT contributed to the data analysis and interpretation. YW, BLevis, AB, and BDT contributed to drafting the manuscript. All authors provided a critical review and approved the final manuscript. AB and BDT are guarantors.

Conflict of interest. Drs Jetté and Patten declare that they received a grant, outside the submitted work, from the University of Calgary Hotchkiss Brain Institute, which was jointly funded by the Institute and Pfizer. Pfizer was the original sponsor of the development of the PHQ-9, which is now in the public domain. Dr Chan is a steering committee member or consultant of Astra Zeneca, Bayer, Lilly, MSD and Pfizer. She has received sponsorships and honorarium for giving lectures and providing consultancy and her affiliated institution has received research grants from these companies. Dr Hegerl declares that within the last three years, he was an advisory board member for Lundbeck and Servier; a consultant for Bayer Pharma; a speaker for Roche Pharma and Servier; and received personal fees from Janssen, all outside the submitted work. Dr Inagaki declares that he has received a grant from Novartis Pharma, and personal fees from Meiji, Mochida, Takeda, Novartis, Yoshitomi, Pfizer, Eisai, Otsuka, MSD, Technomics, and Sumitomo Dainippon, all outside of 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.

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; [2]Department of Psychiatry, McGill University, Montréal, Québec, Canada; [3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; [4]Department of Psychology, McGill University, Montréal, Québec, Canada; [5]Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montreal, Quebec, Canada; [6]Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit, Amsterdam, the Netherlands; [7]Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK; [8]Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA; [9]Library, Concordia University, Montréal, Québec, Canada; [10]Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada; [11]Hotchkiss Brain Institute and O'Brien Institute for Public Health, University of Calgary, Calgary, Alberta, Canada; [12]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; [13]Department of Psychiatry, Makerere University College of Health Sciences, Kampala, Uganda; [14]Department of General Practice and Primary Health Care, University of Auckland, Auckland, New Zealand; [15]Louis and Gabi Weisfeld School of Social Work, Bar Ilan University, Ramat Gan, Israel; [16]Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; [17]Ageing Clinical & Experimental Research Team, Institute of Applied Health Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, Scotland, UK; [18]Department of Medicine, McGill University, Montréal, Québec, Canada; [19]Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA; [20]Centre for Research on Ageing, Health and Wellbeing, Research School of Population Health, The Australian National University, Canberra, Australia; [21]Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Melbourne, Australia; [22]Centre for Brain and Mental Health Research, University of Newcastle, New South Wales, Australia; [23]Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; [24]Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region (SAR), China; [25]Asia Diabetes Foundation, Prince of Wales Hospital, Hong Kong SAR, China; [26]Hong Kong Institute of Diabetes and Obesity, Hong Kong SAR, China; [27]Department of Pediatrics, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA; [28]Department of Psychiatry, University of Rochester Medical Center, New York, USA; [29]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands; [30]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA; [31]Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, Berlin, Germany; [32]Department of Child & Adolescent Psychiatry, Institute of Mental Health, Singapore, Singapore; [33]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore; [34]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore; [35]Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore, Singapore; [36]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; [37]School of Social Work, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; [38]Global and Community Mental Health Research Group, Department of Psychology, Faculty of Social Sciences, University of Macau, Macau Special Administrative Region, China; [39]Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA; [40]City of Minneapolis Health Department, Minneapolis, Minnesota, USA; [41]Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [42]Depression Research Center of the German Depression Foundation and Department of Psychiatry, Psychosomatics and Psychotherapy, Goethe University, Frankfurt, Germany; [43]School of Psychology, University of Queensland, Brisbane, Queensland, Australia; [44]STAR-Stress, Anxiety & Resilience Consultants, Chicago, Illinois, USA; [45]Department of Psychiatry, University of Ioannina, Ioannina, Greece; [46]Department of Psychiatry, Faculty of Medicine, Shimane University, Shimane, Japan; [47]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [48]School of Psychology, University of New South Wales, Sydney, Australia; [49]Neuroscience Research Australia, Sydney, Australia; [50]Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore, Singapore; [51]Department of Psychiatry, Amsterdam UMC, Vrije Universiteit, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands; [52]Department of Psychiatry, Mackay Memorial Hospital, Taipei, Taiwan; [53]Department of Medical Research, Mackay Memorial Hospital, Taipei, Taiwan; [54]Department of Medicine, Mackay Medical

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press

College, Taipei, Taiwan; [55]Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; [56]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [57]Department of Nursing, St. Joseph's College, Standish, Maine, USA; [58]Department of Psychiatry, Faculty of Medicine and Health Sciences, Cancer Resource & Education Centre, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; [59]Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil; [60]Department of Clinical Psychology, Graduate School of Niigata Seiryo University, Niigata, Japan; [61]National Institute of Science and Technology, Translational Medicine, Ribeirão Preto, Brazil; [62]Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, USA; [63]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; [64]Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; [65]Department of Adult Psychiatry, University Hospitals Leuven, Leuven, Belgium; [66]Department of Neurosciences, Katholieke Universiteit Leuven, Leuven, Belgium; [67]Centre for Behavioural Sciences and Mental Health, Italian National Institute of Health, Rome, Italy; [68]Department of Psychiatry and Psychotherapy, University Medical Center Freiburg, Freiburg, Germany; [69]Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburg, Edinburgh, Scotland, UK; [70]Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia; [71]Allina Health, Minneapolis, Minnesota, USA; [72]Centre for Women's Mental Health, Royal Women's Hospital, Parkville, Melbourne, Australia; [73]Melbourne School of Psychological Sciences, University of Melbourne, Australia; [74]School of Medicine and Public Health, University of Newcastle, New South Wales, Newcastle, Australia; [75]IMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, Victoria, Australia; [76]Department of General Practice, Amsterdam Institute for General Practice and Public Health, Amsterdam University Medical Centers, location AMC, Amsterdam, the Netherlands; [77]Monash University, Melbourne, Australia; [78]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA; [79]Department of Medicine, Veterans Affairs Medical Center, San Francisco, California, USA; [80]Department of Medicine, University of California San Francisco, San Francisco, California, USA; [81]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, UK; [82]Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Ogawa-Higashi, Kodaira, Tokyo, Japan; [83]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada; [84]Department of Educational and Counselling Psychology, McGill University, Montréal, Québec, Canada

References

Allaby M (2010) *Screening for Depression: A Report for the UK National Screening Committee* (Revised report). London, United Kingdom: UK National Screening Committee.

American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Arlington, VA: American Psychiatric Publishing.

Barrera TL, Cummings JP, Armento M, Cully JA, Bush Amspoker A, Wilson NL, Mallen MJ, Shrestha S, Kunik ME and Stanley MA (2017) Telephone-delivered cognitive-behavioral therapy for older, rural veterans with depression and anxiety in home-based primary care. *Clinical Gerontologist* **40**, 114–123.

Brugha TS, Bebbington PE and Jenkins R (1999) A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychological Medicine* **29**, 1013–1020.

Canadian Agency for Drugs and Technologies in Health (2016) *PRESS – Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration* (PRESS E&E). Ottawa: CADTH.

Carter G and Spittal MJ (2018) Suicide risk assessment: risk stratification is not accurate enough to be clinically useful and alternative approaches are needed. *Crisis* **39**, 229.

Corson K, Gerrity MS and Dobscha SK (2004) Screening for depression and suicidality in a VA primary care setting: 2 items are better than 1 item. *The American Journal of Managed Care* **10**, 839–845.

Crawford MJ, Thana L, Methuen C, Ghosh P, Stanley SV, Ross J, Gordon F, Blair G and Bajaj P (2011) Impact of screening for risk of suicide: randomised controlled trial. *British Journal of Psychiatry* **198**, 379–384.

Fagerland MW, Lydersen S and Laake P (2014) Recommended tests and confidence intervals for paired binomial proportions. *Statistics in Medicine* **33**, 2850–2875.

Higgins JP and Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539–1558.

Ishihara M, Harel D, Levis B, Levis AW, Riehm KE, Saadat N, Azar M, Rice DB, Sanchez TA, Chiovitti MJ and Cuijpers P (2019). Shortening self-report mental health symptom measures through optimal test assembly methods: development and validation of the Patient Health Questionnaire-Depression-4. *Depression and Anxiety* **36**, 82–92. doi: 10.1002/da.22841.

Kroenke K, Spitzer RL and Williams JBW (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16**, 606–613.

Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT and Mokdad AH (2009) The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* **114**, 163–173.

Levis B, Benedetti A, Riehm KE, Saadat N, Levis AW, Azar M, Rice DB, Chiovitti MJ, Sanchez TA, Cuijpers P and Gilbody S (2018) Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. *British Journal of Psychiatry* **212**, 377–385.

Levis B, Benedetti A, Thombs BD and on behalf of the DEPRESsion Screening Data (DEPRESSD) Collaboration (2019) Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* **365**, l1476.

Moriarty AS, Gilbody S, McMillan D and Manea L (2015) Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *General Hospital Psychiatry* **37**, 567–576.

Nosen E and Woody SR (2008) Chapter 8: diagnostic assessment in research. In McKay D (ed.), *Handbook of Research Methods in Abnormal and Clinical Psychology*. Thousand Oaks: Sage, pp. 109–124.

Razykov I, Ziegelstein R, Whooley M and Thombs BD (2012) The PHQ-9 versus the PHQ-8 – is item 9 useful for assessing suicide risk in coronary artery disease patients? Data from the heart and Soul study. *Journal of Psychosomatic Research* **73**, 163–168.

Rice DB, Kloda LA, Levis B, Qi B, Kingsland E and Thombs BD (2016) Are MEDLINE searches sufficient for systematic reviews and meta-analyses of the diagnostic accuracy of depression screening tools? A review of meta-analyses. *Journal of Psychosomatic Research* **87**, 7–13.

Riley RD, Dodd SR, Craig JV, Thompson JR and Williamson PR (2008) Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine* **27**, 6111–6136.

Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, Farmer A, Jablenski A, Pickens R, Regier DA and Sartorius N (1988) The Composite International Diagnostic Interview: an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* **45**, 1069–1077.

Sheehan DV, Lecrubier Y, Sheehan KH, Janavs J, Weiller E, Keskiner A, Schinka J, Knapp E, Sheehan MF and Dunbar GC (1997) The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry* **12**, 232–241.

Simon GE, Rutter CM, Peterson D, Oliver M, Whiteside U, Operskalski B and Ludman EJ (2013) Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? *Psychiatric Services* **64**, 1195–1202.

Simon GE, Coleman KJ, Rossom RC, Beck A, Oliver M, Johnson E, Whiteside U, Operskalski B, Penfold RB, Shortreed SM and Rutter C (2016) Risk of suicide attempt and suicide death following completion of the Patient Health Questionnaire depression module in community practice. *The Journal of Clinical Psychiatry* **77**, 221–227.

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press

Siu AL and the US Preventive Services Task Force (USPSTF) (2016) Screening for depression in adults: US preventive services task force recommendation statement. *JAMA: The Journal of the American Medical Association* **315**, 380–387.

Suarez L, Beach SR, Moore SV, Mastromauro CA, Januzzi JL, Celano CM, Chang TE and Huffman JC (2015) Use of the Patient Health Questionnaire-9 and a detailed suicide evaluation in determining imminent suicidality in distressed patients with cardiac disease. *Psychosomatics* **56**, 181–189.

Thombs BD, Benedetti A, Kloda LA, Levis B, Nicolau I, Cuijpers P, Gilbody S, Ioannidis JP, McMillan D, Patten SB and Shrier I (2014) The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Systematic Reviews* **3**, 124.

van der Leeden R, Busing FMTA and Meijer E (1997) *Bootstrap Methods For Two-Level Models* (Technical report PRM 97-04). Leiden, The Netherlands: Leiden University, Department of Psychology.

van der Leeden R, Meijer E and Busing FMTA (2008) Chapter 11: resampling multilevel models. In Leeuw J and Meijer E (eds), *Handbook of Multilevel Analysis*. New York: Springer, pp. 401–433.

Walker E and Nowacki AS (2011) Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine* **26**, 192–196.

Walker J, Hansen CH, Butcher I, Sharma N, Wall L, Murray G and Sharpe M (2011) Thoughts of death and suicide reported by cancer patients who endorsed the "suicidal thoughts" item of the PHQ-9 during routine screening for depression. *Psychosomatics* **52**, 424–427.

Wells TS, Horton JL, LeardMann CA, Jacobson IG and Boyko EJ (2013) A comparison of the PRIME-MD PHQ-9 and PHQ-8 in a large military prospective study, the Millennium Cohort Study. *Journal of Affective Disorders* **148**, 77–83.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA and Bossuyt PM (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* **155**, 529–536.

https://doi.org/10.1017/S0033291719001314 Published online by Cambridge University Press