This is a repository copy of *Selective cutoff reporting in studies of the accuracy of the Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale: comparison of results based on published cutoffs versus all cutoffs using individual participant data meta-analysis*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/181466/

Version: Published Version

ORIGINAL ARTICLE

WILEY

Selective cutoff reporting in studies of the accuracy of the Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale: Comparison of results based on published cutoffs versus all cutoffs using individual participant data meta-analysis

Dipika Neupane[1,2] | Brooke Levis[1,2,3] | Parash M. Bhandari[1,2] |
Brett D. Thombs[1,2,4,5,6,7,8] | Andrea Benedetti[2,5,9] |
DEPRESsion Screening Data (DEPRESSD) Collaboration

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

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

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

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

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

[6]Department of Psychology, McGill University, Montréal, Québec, Canada

[7]Department of Educational and Counselling Psychology, McGill University, Montréal, Québec, Canada

[8]Biomedical Ethics Unit, McGill University, Montréal, Québec, Canada

[9]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada

Correspondence
Andrea Benedetti, Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre, 5252 Boulevard de Maisonneuve, Montréal, Quebec H4A 3S5, Canada.
Email: andrea.benedetti@mcgill.ca

Brett D. Thombs, Jewish General Hospital, 4333 Cote Ste. Catherine Rd., Montreal, Quebec H3T 1E4, Canada.
Email: brett.thombs@mcgill.ca

Abstract

Objectives: Selectively reported results from only well-performing cutoffs in diagnostic accuracy studies may bias estimates in meta-analyses. We investigated cutoff reporting patterns for the Patient Health Questionnaire-9 (PHQ-9; standard cutoff 10) and Edinburgh Postnatal Depression Scale (EPDS; no standard cutoff, commonly used 10–13) and compared accuracy estimates based on published cutoffs versus all cutoffs.

Methods: We conducted bivariate random effects meta-analyses using individual participant data to compare accuracy from published versus all cutoffs.

Results: For the PHQ-9 (30 studies, $N = 11{,}773$), published results underestimated sensitivity for cutoffs below 10 (median difference: -0.06) and overestimated for cutoffs above 10 (median difference: 0.07). EPDS (19 studies, $N = 3637$) sensitivity estimates from published results were similar for cutoffs below 10 (median difference: 0.00) but higher for cutoffs above 13 (median difference: 0.14). Specificity estimates from published and all cutoffs were similar for both tools. The mean cutoff of all reported cutoffs in PHQ-9 studies with optimal cutoff below 10 was 8.8 compared to 11.8 for those with optimal cutoffs above 10. Mean for EPDS studies with optimal cutoffs below 10 was 9.9 compared to 11.8 for those with optimal cutoffs greater than 10.

Conclusion: Selective cutoff reporting was more pronounced for the PHQ-9 than EPDS.

KEYWORDS
diagnostic test accuracy, individual participant data meta-analysis, meta-analysis, publication bias, selective cutoff reporting

Sciences and Primary Care, Department of Nursing Science, University Medical Center Utrecht—University Utrecht, Utrecht, the Netherlands; Jaime Delgadillo, Clinical Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, UK; Jesse R. Fann, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA; Nicolas Favez, Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland; Daniel Fung, Department of Developmental Psychiatry, Institute of Mental Health, Singapore; Lluïsa Garcia-Esteve, Perinatal Mental Health Unit CLINIC-BCN. Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain; Bizu Gelaye, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; Felicity Goodyear-Smith, Department of General Practice and Primary Health Care, University of Auckland, Auckland, New Zealand; Thomas Hyphantis, Department of Psychiatry, Faculty of Medicine, University of Ioannina, Ioannina, Greece; Masatoshi Inagaki, Department of Psychiatry, Faculty of Medicine, Shimane University, Shimane, Japan; Khalida Ismail, Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neurosciences, King's College London Weston Education Centre, London, UK; Nathalie Jetté, Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada; Dina Sami Khalifa, Department of Community Medicine, Institute of Health and Society, Faculty of Medicine, University of Oslo, Oslo, Norway; Mohammad E. Khamseh, Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; Jane Kohlhoff, School of Psychiatry, University of New South Wales, Kensington, Australia; Zoltán Kozinszky, Department of Obstetrics and Gynecology, Danderyd Hospital, Stockholm, Sweden; Laima Kusminskas, Private Practice, Hamburg, Germany; Shen-Ing Liu, Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore; Manote Lotrakul, Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; Sonia R. Loureiro, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Bernd Löwe, Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Sherina Mohd Sidik, Cancer Resource & Education Centre, and Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; Sandra Nakić Radoš, Department of Psychology, Catholic University of Croatia, Zagreb, Croatia; Flávia L. Osório, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Susan J. Pawlby, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; Brian W. Pence, Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Tamsen J. Rochat, MRC/Developmental Pathways to Health Research Unit, Faculty of Health Sciences, University of Witwatersrand, South Africa; Alasdair G. Rooney, Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburg, Edinburgh, Scotland, UK; Deborah J. Sharp, Centre for Academic Primary Care, Bristol Medical School, University of Bristol, UK; Lesley Stafford, Centre for Women's Mental Health, Royal Women's Hospital, Parkville, Australia; Kuan-Pin Su, An-Nan Hospital, China Medical University and Mind-Body Interface Laboratory, China Medical University Hospital, Taiwan; Sharon C. Sung, Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore; Meri Tadinac, Department of Psychology, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia; S. Darius Tandon, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; Pavaani Thiagayson, The Institute of Mental Health, Singapore; Annamária Töreki, Department of Emergency, University of Szeged, Hungary; Anna Torres-Giménez, Perinatal Mental Health Unit CLINIC-BCN. Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain; Alyna Turner, School of Medicine and Public Health, University of Newcastle, New South Wales, Newcastle, Australia; Christina M. van der Feltz-Cornelis, Department of Health Sciences, HYMS, University of York, York, UK; Johann M. Vega-Dienstmaier, Facultad de Medicina Alberto Hurtado, Universidad Peruana Cayetano Heredia, Lima, Perú; Paul A. Vöhringer, Department of Psychiatry and Mental Health, Clinical Hospital, Universidad de Chile, Santiago, Chile; Jennifer White, Department of Physiotherapy, School of Primary and Allied Health Care, Monash University, Melbourne, Australia; Mary A. Whooley, Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA; Kirsty Winkley, Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, UK; Mitsuhiko Yamada, Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Ogawa-Higashi, Kodaira, Tokyo, Japan

1 | INTRODUCTION

Selective reporting occurs when authors make decisions regarding publication of study results based on whether or not outcomes are favorable (Kirkham et al., 2010). In accuracy studies of ordinal or continuous tests, selective cutoff reporting occurs when results are published for one or more cutoffs that maximize sensitivity and specificity in a particular study but not for other relevant cutoffs (Levis et al., 2017; Moriarty et al., 2015). Selective cutoff reporting can lead to overestimation of diagnostic accuracy in primary studies and in meta-analyses that synthesize results from primary studies with selectively reported results (Leeflang et al., 2008).

Only one previous study has investigated selective cutoff reporting patterns in test accuracy studies (Levis et al., 2017). That study obtained individual participant data (IPD) from 13 primary studies included in a published meta-analysis (Manea et al., 2012) of the accuracy of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool. Results based on two sets of meta-analysis were compared. First, meta-analyses were conducted where the result at each cutoff was based only on those studies that published results at that cutoff. Second, meta-analyses were conducted based on the IPD; the result at each cutoff was calculated from all studies available regardless of what cutoff was originally published. Sensitivity estimates differed substantially between published and IPD datasets for cutoffs lower and higher than the standard cutoff of 10 (meaning cutoff ≥10) but were similar at the standard cutoff. This was because most studies published results for the standard cutoff, but authors tended to publish results from cutoffs lower or higher than 10 depending on whether the PHQ-9 was relatively poorly sensitive but specific (lower cutoffs published) or highly sensitive but poorly specific (higher cutoffs published) in their dataset.

A cutoff of 10 is used as the standard cutoff for screening for major depression with the PHQ-9 (Gilbody et al., 2007; Kroenke et al., 2001; Kroenke & Spitzer, 2002; Spitzer et al., 1999; Wittkampf et al., 2007) and maximizes combined sensitivity and specificity (Levis et al., 2019), but standard cutoffs are less well-defined for other depression screening tools. Studies of the Edinburgh Postnatal Depression Scale (EPDS), the most commonly used screening tool among women in pregnancy and postpartum (Hewitt et al., 2009; Howard et al., 2014), typically consider cutoffs between 10 and 13 as standard, with 13 being most commonly used (Hewitt et al., 2009; O'Connor et al., 2016). A recent IPD meta-analysis (IPDMA) found that cutoff 11 maximizes combined sensitivity and specificity (Levis et al., 2020).

The degree to which there is an agreed upon standard cutoff for a screening tool may influence selective cutoff reporting. Thus, this study aimed to compare selective cutoff reporting in screening tools with and without a well-defined standard cutoff. We evaluated selective cutoff reporting with a substantially larger set of PHQ-9 studies than was used in the previous study (Levis et al., 2017) and compared results to the EPDS, which does not have a well-defined standard cutoff. Specific objectives were to use IPDMA with the PHQ-9 and EPDS, separately, to (1) compare sensitivity and specificity based on all cutoffs from all primary studies versus data from only cutoffs for which accuracy estimates were published in the primary studies; and (2) explore cutoff reporting patterns with reference to the identified optimal cutoff in each primary study.

2 | METHODS

We analyzed data accrued for IPDMAs on PHQ-9 and EPDS diagnostic accuracy (PROSPERO CRD42014010673, CRD42015024785), and protocols were published for each IPDMA (Thombs et al., 2014, 2015). The protocol for the present study, which was not part of the main IPDMA protocols, was published separately (https://osf.io/vw3bz/). The protocol described only the EPDS analysis, and we subsequently added the PHQ-9 to be able to compare screening tools with and without well-defined standard cutoffs. As this study involved only analysis of previously collected de-identified data and because all included studies were required to have obtained ethics approval and informed consent, the Research Ethics Committee of the Jewish General Hospital determined that ethics approval was not required.

2.1 | Study eligibility

Datasets from articles in any language were eligible for the main IPDMAs if (1) they used the PHQ-9 or EPDS; (2) they included diagnostic classification for current Major Depressive Disorder (MDD) or Major Depressive Episode (MDE) using Diagnostic and Statistical Manual of Mental Disorders (DSM) or International Classification of Diseases (ICD) criteria based on a validated diagnostic interview; (3) the interview and PHQ-9 or EPDS were administered within 2 weeks of each other; (4) participants were ≥18 years and not recruited from school-based settings (PHQ-9) or ≥18 years and pregnant or within 12 months postpartum (EPDS); and (5) participants were not recruited from psychiatric settings or because they had symptoms of depression, since screening is done to identify previously unrecognized cases. Datasets where not all participants were eligible were included if primary data allowed selection of eligible participants.

Many primary studies in the main IPDMA databases that contributed eligible datasets never published estimates of screening accuracy. Thus, for the present study, we restricted analyses to primary studies with publications that included sensitivity and specificity estimates for at least one PHQ-9 or EPDS cutoff for identifying major depression. We excluded studies if the sample size from the published primary study differed by >10% from the sample included in our IPDMA datasets. Sample sizes from original primary studies and the IPDMA databases differed in some cases because, for instance, we excluded participants who were included in the original studies if there were >2 weeks between their index test and reference standard administrations or if they were <18 years old. We also excluded primary studies with publications that reported accuracy

results only for diagnostic classifications broader than major depression (e.g., "any depressive disorder") if the number of cases in the published article and IPDMA datasets differed by >10%.

2.2 | Search strategy and study selection

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations and PsycINFO via OvidSP, and Web of Science via ISI Web of Knowledge from January 1, 2000 to February 7, 2015 (Method S1a) for the PHQ-9 and from inception to June 10, 2016 (Method S1b) for the EPDS, using peer-reviewed search strategies (McGowan et al., 2016). We also reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS) for de-duplication and then into DistillerSR (Evidence Partners).

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

2.3 | Data contribution, extraction, and synthesis

Authors of eligible datasets were emailed invitations to contribute de-identified primary data at least three times, as necessary. If there was no response, we emailed co-authors and attempted phone contact. For each study, we compared published results with results from raw datasets and resolved any discrepancies in consultation with primary study investigators. For defining major depression, we considered MDD or MDE based on DSM or ICD. If more than one was reported, we prioritized MDE over MDD and DSM over ICD. For studies with multiple time points, we included data from only the time point with the most participants. To facilitate comparison between published results and IPDMA results, we applied sampling weights in the IPDMA only when accuracy results reported in the original published study were calculated using weights.

We determined whether included primary studies cited the Standards for Reporting of Diagnostic Test Accuracy (STARD) guideline in the publication or not (Bossuyt et al., 2003).

2.4 | Statistical analyses

We replicated the statistical analyses used in the previous study of selective cutoff reporting with the PHQ-9 (Levis et al., 2017). We estimated sensitivity and specificity from cutoffs up to 5 points below and above cutoffs used as standard (PHQ-9 cutoff 10, range 5–15; EPDS cutoffs 10–13, range 5–18). We compared meta-analyses results from data using only cutoffs for which accuracy estimates were

published in the primary studies (the *published dataset*) and using data from all cutoffs from all studies (the *full dataset*).

For both sets of meta-analyses, for each cutoff, bivariate random-effects models were estimated via Gauss-Hermite quadrature (Riley et al., 2008). This approach models sensitivity and specificity simultaneously, accounting for the inherent correlation between them and the precision of estimates within studies.

2.4.1 | Differences in sensitivity and specificity estimates using *published* versus *full datasets*

In order to examine differences in results produced by meta-analyses based on *published* and *full datasets*, we constructed separate pooled receiver operator characteristic (ROC) curves. In addition, 95% confidence intervals for the differences in sensitivity and specificity at each cutoff were constructed via bootstrap (Van der Leeden et al., 1997, 2008) resampling at the study and subject level with 1000 iterations for each cutoff. We calculated the median absolute difference in estimated sensitivity and specificity across evaluated cutoffs.

2.4.2 | Reporting patterns

We assessed whether primary studies tended to preferentially report low or high cutoffs depending on the study's sample-specific optimal cutoff. For each primary study, we identified the optimal cutoff that the authors explicitly described as optimal or using a similar term. If the authors did not identify an optimal cutoff, we used the cutoff that maximized Youden's J (sensitivity + specificity−1) (Youden, 1950). For each study, we plotted the optimal cutoff, along with all other cutoffs for which results were published. We noted whether the reported cutoffs tended to be low or high compared to the standard cutoff (PHQ-9 10) or set of commonly used cutoffs (EPDS 10–13). For studies with optimal cutoffs below and above the standard or commonly used cutoffs, separately, we calculated the mean of the cutoffs reported.

3 | RESULTS

3.1 | Identification of eligible studies

3.1.1 | Patient Health Questionnaire-9

Of 58 studies included in the main IPDMA (Levis et al., 2019), 28 were excluded from the present study because they did not publish diagnostic accuracy results for any PHQ-9 cutoffs or because the number of participants or major depression cases in the IPD dataset differed by >10% from the published studies or could not be determined (Figure S1a; Tables S1a and S2a). The final dataset included 30 studies (N total: 11,773; N major depression: 1587 [13%]; Table S3a)

that compared the PHQ-9 with a validated diagnostic interview (Mini Neurospsychiatric Diagnostic Interview, Structured Clinical Interview for DSM Disorders, Composite International Diagnostic Interview, Clinical Interview Schedule Revised, Schedules for Clinical Assessment in Neuropsychiatry or Computerized Diagnostic Interview Schedule). Of the 30 included studies, 7 reported only a single cutoff and 23 reported more than one cutoff. Of the 23 with multiple cutoffs reported, 18 identified an optimal cutoff in the published study; of those, 16 (89%) were described as based on Youden's J (N: 8) or equivalent to Youden's calculated from published cutoffs but did not have an explanation (N: 8). Among the 30 studies, only two cited the STARD reporting guideline (Arroll et al., 2010; Sherina et al., 2012).

3.1.2 | Edinburgh Postnatal Depression Scale

Of 49 studies in the original IPDMA dataset (Levis et al., 2020), 30 studies were not eligible and thus excluded from the present study (Figure S1b; Tables S1b and S2b). Thus, 19 unique studies (N total: 3637, N major depression: 531 [15%]) were included (Table S3b), which compared the EPDS with a validated diagnostic interview including Mini Neuropsychiatric Diagnostic Interview, Structured Clinical Interview for DSM Disorders, Clinical Interview Schedule and Diagnostic Interview of Genetic Studies. Of the 14 studies that reported more than one cutoff, 13 identified an optimal cutoff; of those 10 (77%) were based on Youden's J (N: 2) or did not have an explanation but matched what would have been obtained using Youden's J calculated from published cutoffs (N: 8). None of the studies cited STARD.

3.2 | Differences in sensitivity and specificity estimates based on *published* versus *full datasets*

Table 1 shows sensitivity and specificity for the PHQ-9 and EPDS at each cutoff for the *published* and *full datasets* with the ROC plots in Figures 1 and 2.

3.2.1 | Patient Health Questionnaire-9

The difference between estimated sensitivity (*published−full dataset*) ranged from −0.09 to 0.10 (median: 0.06; Table 2). For cutoffs below 10, estimated sensitivity was lower for the *published dataset* (−0.02 to −0.09; median: −0.06) with 95% CIs including zero but inclining more towards negative, whereas estimated specificity was higher (0.01 to 0.14; median: 0.03) with 95% CIs including zero. For the standard cutoff 10, the differences in sensitivity and specificity were −0.01 (95% CI: −0.05, 0.01), and 0.01 (95% CI: 0.00, 0.04), respectively. For cutoffs above 10, estimated sensitivity was higher for the *published dataset* (0.00 to 0.10; median: 0.07) with 95% CIs including zero but inclining more towards positive, and estimated

specificity was similar (0.00 to 0.02; median: 0.01) with 95% CIs including zero.

3.2.2 | Edinburgh Postnatal Depression Scale

The difference between estimated sensitivity ranged from −0.02 to 0.20 (median: 0.03) with all 95% CIs including zero (Table 2). For cutoffs below 10, estimated sensitivity (−0.02 to 0.01; median: 0.00), and estimated specificity (-0.01 to 0.02; median: 0.01) were similar for the *published and full datasets*. For cutoffs of 10 to 13, estimated sensitivity differed by 0.02 to 0.03 (median: 0.03), and estimated specificity differed by -0.02 to 0.00 (median: -0.02). For cutoffs above 13, estimated sensitivity was higher for the *published dataset* (0.08 to 0.20; median: 0.14), and estimated specificity was similar or lower (-0.08 to 0.00; median: 0.00).

3.3 | Reporting patterns

3.3.1 | Patient Health Questionnaire-9

Figure 3 shows the pattern of reporting with respect to optimal cutoffs for included PHQ-9 studies; 9 studies had optimal cutoffs below 10, 14 equal to 10, 6 greater than 10 and 1 study had optimal cutoffs of both 10 and 12. Studies for which the PHQ-9 was poorly sensitive at the cutoff 10 (sensitivity: 0.27–0.74) (Arroll et al., 2010; Inagaki et al., 2013; Lambert et al., 2015; Lotrakul et al., 2008; Pence et al., 2012; Thombs et al., 2008; Stafford et al., 2007; Sung et al., 2013; Turner et al., 2012) had optimal cutoffs that were below 10. These studies tended to report more cutoffs below 10 than above 10 (mean of reported cutoffs: 8.8). Studies for which the PHQ-9 was highly sensitive at cutoff 10 (sensitivity: 0.85–1.00) (Bombardier et al., 2012; Delgadillo et al., 2011; Fann et al., 2005; Khamseh et al., 2011; Lowe et al., 2004; Twist et al., 2013) had optimal cutoffs that were greater than 10. These studies tended to report more cutoffs above 10 than below 10 (mean of reported cutoffs: 11.8).

3.3.2 | Edinburgh Postnatal Depression Scale

Figure 4 shows the pattern of reporting cutoffs for the EPDS; 5 studies had optimal cutoffs below 10, 13 between 10 and 13, and 1 greater than 13. Studies for which the EPDS was poorly sensitive at cutoff 10 (sensitivity: 0.43–0.73) (Bakare et al., 2014; Chaudron et al., 2010; Radoš et al., 2013; Thiagayson et al., 2013; Toreki et al., 2013) had optimal cutoffs that were less than 10 (mean of reported cutoffs: 9.9). Studies for which EPDS was highly sensitive at cutoff 10 (sensitivity: 0.82–1.00) (Alvarado et al., 2015; Beck & Gable, 2001; Bunevicius et al., 2009; Couto et al., 2015; Garcia-Esteve et al., 2003; Khalifa et al., 2015; Phillips et al., 2009; Rochat et al., 2013; Su et al., 2007; Tandon et al., 2012; Toreki et al., 2014; Vega-Dienstmaier et al., 2002) had optimal cutoffs greater than 10.

TABLE 1 Comparison of accuracy results from IPDMA of PHQ-9 and EPDS with the *published dataset* only versus the *full dataset*

PHQ-9

| | **Published dataset** | | | | | | | **Full dataset 30 studies; N = 11 773; MD cases = 1587** | | | |
Cutoff	No. of studies	No. of participants	No of MD cases	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	5	1663	367	0.91	0.86, 0.94	0.68	0.55, 0.79	0.97	0.94, 0.98	0.54	0.48, 0.60
6	6	2193	377	0.87	0.77, 0.93	0.72	0.61, 0.82	0.96	0.92, 0.97	0.62	0.56, 0.68
7	6	2050	438	0.87	0.75, 0.93	0.72	0.60, 0.81	0.94	0.90, 0.97	0.69	0.63, 0.74
8	12	5798	720	0.87	0.78, 0.92	0.77	0.70, 0.82	0.92	0.87, 0.95	0.75	0.70, 0.79
9	14	5283	766	0.85	0.76, 0.91	0.81	0.75, 0.85	0.87	0.81, 0.91	0.80	0.76, 0.84
10	26	10 593	1378	0.82	0.74, 0.88	0.86	0.83, 0.89	0.83	0.76, 0.88	0.85	0.81, 0.88
11	15	5292	767	0.83	0.72, 0.91	0.88	0.83, 0.92	0.76	0.69, 0.82	0.88	0.85, 0.91
12	16	6188	832	0.73	0.63, 0.81	0.91	0.87, 0.94	0.69	0.62, 0.75	0.91	0.88, 0.93
13	9	2104	455	0.70	0.59, 0.79	0.95	0.87, 0.98	0.60	0.54, 0.67	0.93	0.91, 0.95
14	5	1231	277	0.63	0.47, 0.76	0.96	0.89, 0.99	0.54	0.47, 0.61	0.95	0.93, 0.96
15	6	3546	374	0.47	0.37, 0.59	0.97	0.97, 0.98	0.47	0.40, 0.54	0.96	0.95, 0.97

EPDS

| | **Published dataset** | | | | | | | **Full dataset 19 studies; N = 3637; MD cases = 531** | | | |
Cutoff	No. of studies	No. of participants	No. of MD cases	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	4	830	52	0.98	0.84, 1.00	0.38	0.18, 0.62	0.98	0.95, 0.99	0.36	0.29, 0.43
6	4	830	52	0.98	0.86, 1.00	0.46	0.23, 0.70	0.97	0.93, 0.98	0.45	0.37, 0.53
7	7	1413	122	0.93	0.84, 0.97	0.56	0.41, 0.70	0.94	0.89, 0.97	0.55	0.47, 0.62
8	9	1920	194	0.92	0.80, 0.97	0.62	0.47, 0.74	0.91	0.85, 0.94	0.63	0.55, 0.71
9	13	2807	342	0.85	0.78, 0.91	0.72	0.63, 0.80	0.87	0.81, 0.91	0.71	0.63, 0.78
10	11	2215	210	0.84	0.73, 0.91	0.78	0.68, 0.85	0.82	0.76, 0.87	0.79	0.72, 0.84
11	13	2462	277	0.83	0.72, 0.90	0.83	0.76, 0.89	0.80	0.72, 0.86	0.85	0.79, 0.90
12	12	2373	252	0.75	0.60, 0.86	0.87	0.80, 0.92	0.72	0.63, 0.80	0.89	0.84, 0.92
13	17	3032	447	0.68	0.59, 0.76	0.93	0.89, 0.96	0.65	0.56, 0.74	0.93	0.89, 0.95
14	9	1950	184	0.66	0.54, 0.76	0.95	0.89, 0.98	0.58	0.49, 0.67	0.95	0.92, 0.97
15	6	1286	131	0.65	0.55, 0.73	0.96	0.90, 0.98	0.50	0.43, 0.58	0.96	0.94, 0.98
16	3	682	65	0.61	0.47, 0.73	0.98	0.78, 1.00	0.41	0.35, 0.49	0.98	0.96, 0.99
17[a]	1	306	19	0.47	0.25, 0.71	0.91	0.87, 0.94	0.33	0.27, 0.41	0.99	0.97, 0.99
18[a]	1	306	19	0.37	0.17, 0.61	0.95	0.92, 0.97	0.26	0.21, 0.33	0.99	0.98, 1.00

Abbreviations: CI, Confidence Interval; EPDS, Edinburgh Postnatal Depression Scale; IPDMA, Individual Participant Data Meta-analysis; MD, Major Depression.

[a]For these cutoffs, one sample proportion test with continuity correction was used to estimate sensitivity and specificity and confidence intervals.

These studies tended to report more cutoffs above 10 than below 10 (mean of reported cutoffs: 11.8). All of these studies had optimal cutoffs between 10 and 13 with one exception, a study reported accuracy only for cutoff 13 even though sensitivity was low at this cutoff (sensitivity: 0.35) (Pawlby et al., 2008).

4 | DISCUSSION

We compared bias in accuracy and selective cutoff reporting between the PHQ-9, which has a clearly defined standard cutoff and the EPDS, which does not have a clearly defined standard cutoff,
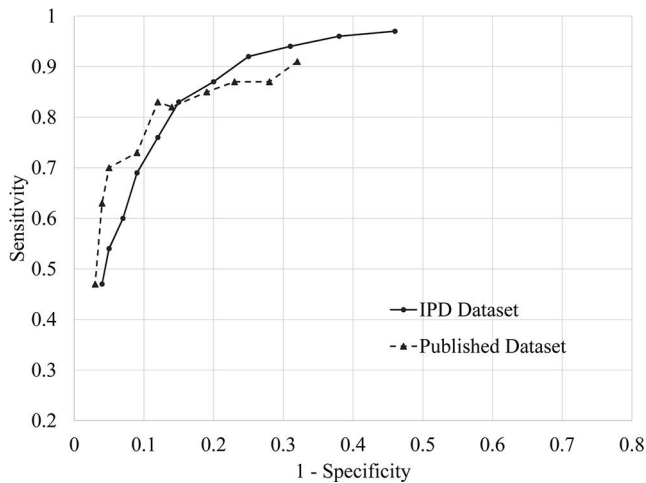
FIGURE 1   Receiver operating characteristic (ROC) curves plot for the diagnostic accuracy of Patient Health Questionnaire-9 (PHQ-9). The points in the ROC curves indicate each of the PHQ-9 cutoffs between 5 (right) and 15 (left)
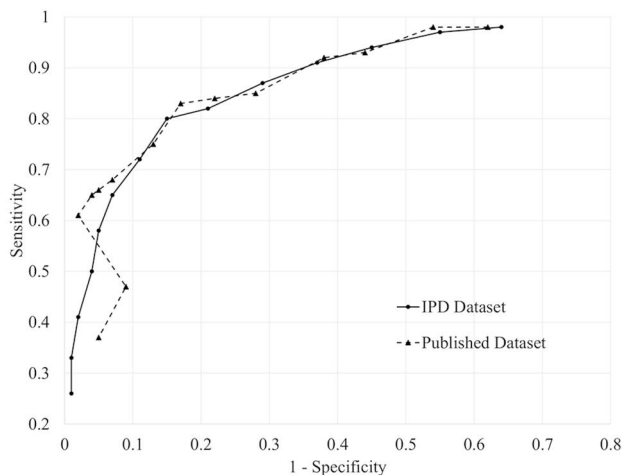

FIGURE 2   Receiver operating characteristic (ROC) curves plot for the diagnostic accuracy of Edinburgh Postnatal Depression Scale (EPDS). The points in the ROC curves indicate each of the EPDS cutoffs between 5 (right) and 18 (left)

using IPD. Selective cutoff reporting was more pronounced for the PHQ-9, and bias in estimated accuracy of published cutoffs compared to all cutoffs was similarly greater for the PHQ-9.

For the PHQ-9, compared to meta-analysis of the *full dataset*, which included results for all relevant cutoffs for all included studies, specificity estimates using the *published dataset*, which included results from published cutoffs only, were similar; however, sensitivity was underestimated in the *published dataset* for cutoffs below 10, similar for the standard cutoff 10, and overestimated for cutoffs above 10. The cutoff reporting pattern in primary studies explains this pattern of under and overestimation of sensitivity. Studies in which the PHQ-9 was poorly sensitive but more specific identified cutoffs below 10 as optimal and reported more cutoffs below 10,

whereas studies in which the PHQ-9 was highly sensitive but less specific identified cutoff above 10 as optimal and reported more cutoffs above 10.

For the EPDS, compared to the *full dataset*, specificity estimates using the *published dataset* was similar across all cutoffs; however, sensitivity estimates were similar for cutoffs below 10 and for the most commonly reported cutoffs 10–13, but overestimated for cutoffs above 13. Unlike the PHQ-9, only primary studies in which the EPDS was highly sensitive at cutoff 10 reported more cutoffs above 10. Studies with poor sensitivity that reported optimal cutoffs below 10 reported results from cutoffs above 10 more often than comparable studies with the PHQ-9. This may be because the PHQ-9 has a single standard cutoff of 10, whereas for the EPDS it is an expectation that results for commonly used cutoffs of 10–13 are reported.

The 2001 PHQ-9 validation study, which included only 41 major depression cases, identified 10 as the standard cutoff (Kroenke et al., 2001; Spitzer et al., 1999). Similarly, the 1987 EPDS validation study, which included only 24 definite or probable major depression cases, suggested that cutoffs of 10 or 13 could be used (Cox et al., 1987). Consequently, most PHQ-9 studies report accuracy for cutoff 10, but selectively reported accuracy for cutoffs other than 10 depending upon the sensitivity at cutoff 10 (Levis et al., 2017; Moriarty et al., 2015). In the absence of a single standard cutoff, EPDS studies often report a range of cutoffs from 10 to 13 (Hewitt et al., 2009; O'Connor et al., 2016).

Only one previous study, an IPDMA with 13 studies (4589 participants, 1037 major depression cases), has examined selective cutoff reporting in screening instruments (for the PHQ-9) (Levis et al., 2017). We replicated the analysis with much larger sample (30 studies; 11,773 participants; 1587 cases) and found that although the reporting patterns were similar, the magnitude of bias was lower in the present study. In the previous study, when the cutoff increased from 9 to 10 and 10 to 11, the sensitivity also increased markedly, an impossible finding if all data are analyzed. In the present study, the sensitivity increased when cutoff increased from 10 to 11, but the increment was minimal. The reduction in the magnitude of bias due to selective reporting compared to the previous study may be due to improved reporting practices over time. This could, however, also be a result of differences in inclusion criteria in the two studies. Of the 13 primary studies included in the previous study, six were excluded from the present study for one of the following reasons: selecting sample for existing distress, mental health diagnosis or from psychiatric settings; having >10% difference in sample size or MDD cases between IPD and published dataset; or administering the PHQ-9 and diagnostic interview more than 2 weeks apart.

Primary studies are often carried out to identify optimal cutoffs and explore accuracy of a screening tool in a specific population; regardless, the full range of cutoffs should be reported. According to STARD reporting guidelines, diagnostic accuracy estimates and precision, as well as the cross tabulation of the index test and the reference standard should be reported (Bossuyt et al., 2015). The guideline should also recommend reporting accuracy estimates for all

TABLE 2 Differences in estimated sensitivity and specificity using the *published dataset* only versus the *full dataset* for PHQ-9 and EPDS

PHQ-9						
% of participants included in published results for each cutoff			**Differences in estimates using *published dataset* versus *full dataset* (published - full)**			
			Sensitivity		**Specificity**	
Cutoff	**% participants**	**% MD cases**	**Estimated difference**	**Bootstrap 95% CI**	**Estimated difference**	**Bootstrap 95% CI**
5	14	23	−0.06	−0.13, 0.00	0.14	0.02, 0.26
6	19	24	−0.09	−0.18, −0.01	0.10	0.00, 0.20
7	17	28	−0.07	−0.20, 0.00	0.03	−0.09, 0.15
8	49	45	−0.05	−0.14, 0.02	0.02	−0.03, 0.08
9	45	48	−0.02	−0.11, 0.05	0.01	−0.04, 0.05
10	90	87	−0.01	−0.05, 0.01	0.01	0.00, 0.04
11	45	48	0.07	0.00, 0.13	0.00	−0.03, 0.03
12	53	52	0.04	−0.03, 0.09	0.00	−0.02, 0.03
13	18	29	0.10	−0.02, 0.20	0.02	−0.04, 0.05
14	10	17	0.09	−0.07, 0.23	0.01	−0.04, 0.04
15	30	24	0.00	−0.12, 0.13	0.01	0.00, 0.03
EPDS						
% of participants included in published results for each cutoff			**Differences in estimates using *published dataset* versus *full dataset* (published−full)**			
			Sensitivity		**Specificity**	
Cutoff	**% participants**	**% MD cases**	**Estimate difference**	**Bootstrap 95% CI**	**Estimate difference**	**Bootstrap 95% CI**
5	23	10	0.00	−0.06, 0.04	0.02	−0.16, 0.21
6	23	10	0.01	−0.04, 0.05	0.01	−0.19, 0.21
7	39	23	−0.01	−0.10, 0.07	0.01	−0.12, 0.15
8	53	37	0.01	−0.09, 0.08	−0.01	−0.13, 0.10
9	77	64	−0.02	−0.08, 0.06	0.01	−0.06, 0.08
10	61	40	0.02	−0.11, 0.10	−0.01	−0.09, 0.06
11	68	52	0.03	−0.06, 0.11	−0.02	−0.08, 0.03
12	65	47	0.03	−0.14, 0.15	−0.02	−0.09, 0.03
13	83	84	0.03	−0.03, 0.10	0.00	−0.02, 0.01
14	54	35	0.08	−0.11, 0.21	0.00	−0.06, 0.03
15	35	25	0.15	0.00, 0.32	0.00	−0.08, 0.03
16	19	12	0.20	−0.03, 0.39	0.00	−0.08, 0.03
17	8	4	0.14	−0.09, 0.37	−0.08	−0.11, −0.04
18	8	4	0.11	−0.12, 0.34	−0.04	−0.07, −0.01

Note: For PHQ-9, 15 iterations (1.5%) that did not produce difference estimates were removed prior to determining the bootstrap CI.

For EPDS, 284 iterations (28.4%) for cutoffs 5-6, 32 iterations (3.2%) for cutoffs 7-15 and 275 iterations (27.5%) for cutoff 16 that did not produce difference estimates were removed prior to determining bootstrap CIs. Only 1 study published EPDS cutoffs 17 and 18, so only participant level resampling was done for published dataset.

Abbreviations: CI, Confidence Interval, EPDS: Edinburgh Postnatal Depression Scale, PHQ-9, Patient Health Questionnaire-9.

Author	Published cutoffs for PHQ-9											No. of published cutoff for each study	Mean of reported cutoffs	Sensitivity at cutoff 10	Specificity at cutoff 10
	5	6	7	8	9	10	11	12	13	14	15				
Inagaki, 2013	O											10	8.50	0.55	0.98
Stafford, 2007		O										3	7.00	0.54	0.91
Sung, 2013		O										1	6.00	0.67	0.91
Thombs, 2008		O										6	5.50	0.54	0.90
Pence, 2012				O								3	10.00	0.27	0.94
Arrol, 2010				O								4	11.25	0.74	0.91
Turner, 2012					O							3	8.67	0.69	0.78
Lambert, 2015					O							4	11.80	0.71	0.82
Lotrakul, 2008					O							10	10.50	0.74	0.85
Gelaye, 2014						O						3	10.00	0.53	0.78
Gjerdingen, 2009						O						1	10.00	0.74	0.91
Mohd Sidik, 2012						O						1	10.00	0.77	0.87
Rooney, 2013						O						4	9.50	0.79	0.86
de Man-van Ginkel, 2012						O						1	10.00	0.80	0.78
Cholera, 2014						O						3	10.0	0.81	0.83
Hyphantis, 2011						O						11	9.50	0.81	0.87
Amoozegar, 2017						O						6	12.50	0.82	0.79
Richardson, 2010						O						6	9.50	0.82	0.86
Liu, 2011						O						3	10.00	0.86	0.94
Akena, 2013						O						6	10.50	0.91	0.89
Vöhringer, 2013						O						1	10.00	0.93	0.77
Chagas, 2013						O						4	9.50	1.00	0.83
Osório, 2009						O						6	15.50	1.00	0.98
van Steenbergen-Weijenburg, 2010						O		O				5	10.00	0.92	0.65
Bombardier, 2012							O					4	10.50	1.00	0.80
Fann, 2005								O				2	11.00	0.88	0.90
Delgadillo, 2011								O				1	12.00	0.94	0.42
Löwe, 2004								O				3	12.00	0.97	0.76
Twist, 2013								O				5	12.00	0.98	0.64
Khamseh, 2011									O			1	13.00	0.85	0.66
No. of studies that published each cutoff	5	6	6	12	14	26	15	16	9	5	6				

F I G U R E   3 Pattern of cutoff reporting for PHQ-9 studies. Cells shaded in gray represent cutoff points for which diagnostic accuracy results are reported in the primary studies. "O" represents the optimal cutoff for PHQ-9 explicitly stated in the studies except for Inagaki et al. (2013), Pence et al. (2012), Arroll (2010), Cholera (2014), Amoozegar (2017), which did not identify an optimal cutoff. For those, Youden's J optimal was calculated from published accuracies. For Gjerdingen (2009) and Vöhringer (2013), only one cutoff was reported without stating whether it was optimal or not. van Steenbergen-Weijenburg 2010 reported 10 and 12 as optimal cutoffs. Studies that reported accuracies for cutoffs beyond presented in the table: Inagaki et al. (2013) reported the accuracy for cutoffs 4–13, Thombs (2008) reported the accuracy for cutoffs 1–10, Lambert et al. (2015) reported the accuracy for cutoffs 5, 9, 10, 15, 20, Hyphantis (2011) reported the accuracy for cutoffs 4–16, Osorio (2009) reported the accuracy for cutoffs 10–21. All the reported cutoffs were included while calculating the mean of reported cutoffs though they are not shown in the figure

relevant cutoffs for ordinal index tests. Citation of the STARD guideline, however, was not common; only 2 of 49 PHQ-9 and EPDS studies (Arroll et al., 2010; Sherina et al., 2012) cited it. When data are missing from some cutoffs in primary studies, conventional meta-analyses based on published cutoffs only may result in biased accuracy estimates. Accuracy estimates can be corrected in meta-analyses using modelling techniques (Benedetti et al., 2020) or by doing IPDMA, which has some advantages, but is highly resource intensive (Cochrane methods: IPD meta-analysis, 2020; Ioannidis et al., 2002; Riley et al., 2010; Stewart & Tierney, 2002).

The major strength of this study is that we compared two depression screening instruments with different characteristics using IPDMA. We explored how the presence of a clearly defined standard cutoff versus the absence of such a standard may be associated with bias in accuracy. A potential limitation is that we calculated the optimal cutoff based on Youden's J for the studies not specifying an optimal cutoff. Those studies may not have considered the cutoff that maximized Youden's J as optimal. However, Youden's J appears to be the most typical method of identifying optimal cutoff thresholds for depression screening measures. In the present study, 16 of 18 (89%)

Author	Published cutoffs for EPDS														No. of published cutoff for each study	Mean of reported cutoffs	Sensitivity at cutoff 10	Specificity at cutoff 10
	5	6	7	8	9	10	11	12	13	14	15	16	17	18				
Töreki, 2013					O										10	9.50	0.43	0.93
Nakić Radoš, 2013					O										8	10.50	0.60	0.82
Bakare, 2014					O										1	9.00	0.66	0.89
Chaudron, 2010					O										2	11.00	0.73	0.84
Thiagayson, 2013					O										6	9.50	0.73	0.74
Tissot, 2015						O									5	11.00	0.50	0.75
Couto, 2015							O								7	11.00	0.86	0.68
Tandon, 2012							O								2	12.00	0.92	0.81
Garcia-Esteve, 2003							O								8	11.50	1.00	0.59
Philips, 2009								O							3	12.00	0.88	0.66
Khalifa, 2015								O							11	8.00	0.89	0.68
Bunevicius, 2009								O							7	12.00	0.92	0.87
Töreki, 2014									O						12	10.50	1.00	0.91
Pawlby, 2008									O						1	13.00	0.61	0.94
Alvarado, 2015									O						10	11.50	0.82	0.82
Beck, 2001									O						1	13.00	0.83	0.86
Su, 2007									O						1	13.00	0.91	0.70
Rochat, 2013									O						1	13.00	0.94	0.50
Vega-Dienstmaier, 2002										O					14	13.50	0.89	0.45
No. of studies that published each cutoff	4	4	7	9	13	11	13	12	17	9	6	3	1	1				

FIGURE 4 Pattern of cutoff reporting for EPDS studies. Cells shaded in gray represent cutoff points for which diagnostic accuracy results are reported in the primary studies. "O" represents the optimal cutoff for EPDS explicitly stated in the studies except for Philips (2009), which did not identify an optimal cutoff. For Philips 2009, Youden's J optimal was calculated from published accuracies. For Bakare et al. (2014), Pawlby et al. (2008), Beck 2001 only one cutoff was reported without stating whether it was optimal or not. Studies that reported accuracies for cutoffs beyond presented in the table: Khalifa et al. (2015) reported accuracy for cutoffs 1–15, Vega-Dienstmaier et al. (2002) reported the accuracy for cutoffs 1–26. All the reported cutoffs were included while calculating the mean of reported cutoffs though they are not shown in the figure

PHQ-9 studies and 10 of 13 (77%) EPDS studies with multiple reported cutoffs that identified an optimal cutoff used Youden's J or identified an optimal cutoff that was equivalent to the Youden's J optimal cutoff. Another possible limitation is that we examined primary studies regardless of the reference standard that was used in each study. We have previously shown that different types of diagnostic interviews perform differently (Wu et al., 2021). We do not believe, however, that the reference standard used would have likely influenced decisions about which cutoffs to report in primary studies.

When studies appeared to report cutoffs selectively depending upon the sensitivity at the standard cutoff, synthesis of accuracy results from published cutoffs led to underestimation of sensitivity below the standard cutoff and overestimation of sensitivity above the standard cutoff. This phenomenon appears to be diluted for EPDS when the standard cutoff is not clearly defined and there is a range of commonly used and reported cutoffs, because the primary studies tend to report a range of cutoffs around the true optimal cutoff. To reduce bias in evidence syntheses, researchers conducting primary studies should report accuracy estimates or a contingency table for all relevant cutoffs, or make their primary data available. Researchers who conduct meta-analyses should use modelling approaches to overcome possible biases from selective cutoff reporting or should use an IPDMA approach.

CONFLICT OF INTEREST

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: Dr. Tonelli declares that he has received a grant from Merck Canada, outside the submitted work. Dr. Vigod declares that she receives royalties from UpToDate, outside the submitted work. Dr. Beck declares that she receives royalties for her Postpartum Depression Screening Scale published by Western Psychological Services. 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. Dr. Ismail declares that she has received honorarium for speaker fees for

educational lectures for Sanofi, Sunovion, Janssen and Novo Nordisk. 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.

AUTHOR CONTRIBUTIONS

Dipika Neupane, Brooke Levis, Parash Mani Bhandari, Jill T. Boruff, Pim Cuijpers, Simon Gilbody, John P. A. Ioannidis, Lorie A. Kloda, Dean McMillan, Scott B. Patten, Ian Shrier, Roy C. Ziegelstein, Liane Comeau, Nicholas D. Mitchell, Marcello Tonelli, Simone N. Vigod, Brett D. Thombs, Andrea Benedetti contributed to conception and design of this study. Jill T. Boruff and Lorie A. Kloda designed and performed database searches for this study. Dickens H. Akena, Rubén Alvarado, Bruce Arroll, Muideen O. Bakare, Hamid R. Baradaran, Cheryl Tatano Beck, Charles H. Bombardier, Adomas Bunevicius, Gregory Carter, Marcos H. Chagas, Linda H. Chaudron, Rushina Cholera, Kerrie Clover, Yeates Conwell, Tiago Castro e Couto, Janneke M. de Man-van Ginkel, Jaime Delgadillo, Jesse R. Fann, Nicolas Favez, Daniel Fung, Lluïsa Garcia-Esteve, Bizu Gelaye, Felicity Goodyear-Smith, Thomas Hyphantis, Masatoshi Inagaki, Khalida Ismail, Nathalie Jetté, Dina Sami Khalifa, Mohammad E. Khamseh, Jane Kohlhoff, Zoltán Kozinszky, Laima Kusminskas, Shen-Ing Liu, Manote Lotrakul, Sonia R. Loureiro, Bernd Löwe, Sherina Mohd Sidik, Sandra Nakić Radoš, Flávia L. Osório, Susan J. Pawlby, Brian W. Pence, Tamsen J. Rochat, Alasdair G. Rooney, Deborah J. Sharp, Lesley Stafford, Kuan-Pin Su, Sharon C. Sung, Meri Tadinac, S. Darius Tandon, Pavaani Thiagayson, Annamária Töreki, Anna Torres-Giménez, Alyna Turner, Christina M. van der Feltz-Cornelis, Johann M. Vega-Dienstmaier, Paul A. Vöhringer, Jennifer White, Mary A. Whooley, Kirsty Winkley, Mitsuhiko Yamada contributed primary dataset to this study. Dipika Neupane, Brooke Levis, Parash Mani Bhandari, Ying Sun, Chen He, Yin Wu, Ankur Krishnan, Zelalem Negeri, Mahrukh Imran, Danielle B. Rice, Kira E. Riehm, Nazanin Saadat, Marleine Azar, Tatiana A. Sanchez, Matthew J. Chiovitti and Alexander W. Levis contributed to data extraction and coding for the meta-analysis. Dipika Neupane, Brooke Levis, Parash Mani Bhandari, Brett D. Thombs and Andrea Benedetti contributed to data analysis and interpretation. Dipika Neupane, Brooke Levis, Parash Mani Bhandari, Brett D. Thombs and Andrea Benedetti contributed to drafting the manuscript. All authors provided a critical review and approved the final manuscript. Brett D. Thombs and Andrea Benedetti are the guarantors; they had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses.

DATA AVAILABILITY STATEMENT

Statistical codes and dataset used in the individual participant data meta-analysis can be requested from the corresponding authors, Drs. Andrea Benedetti and Brett D. Thombs.

ORCID

Brett D. Thombs (iD) https://orcid.org/0000-0002-5644-8432
Andrea Benedetti (iD) https://orcid.org/0000-0002-8314-9497

REFERENCES

Alvarado, R., Jadresic, E., Guajardo, V., & Rojas, G. (2015). First validation of a Spanish-translated version of the Edinburgh Postnatal Depression Scale (EPDS) for use in pregnant women. A Chilean study. *Archives of Women's Mental Health*, 18(4), 607–612. https://doi.org/10.1007/s00737-014-0466-z

Amoozegar, F., Patten, S. B., Becker, W. J., Bulloch, A. G. M., Fiest, K. M., Davenport, W. J., Carroll, C. R., & Jette, N. (2017). The prevalence of depression and the accuracy of depression screening tools in migraine patients. *General Hospital Psychiatry*, 48, 25–31. https://doi.org/10.1016/j.genhosppsych.2017.06.006

Arroll, B., Goodyear-Smith, F., Crengle, S., Gunn, J., Kerse, N., Fishman, T., Falloon, K., & Hatcher, S. (2010). Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *The Annals of Family Medicine*, 8(4), 348–353. https://doi.org/10.1370/afm.1139

Bakare, M. O., Okoye, J. O., & Obindo, J. T. (2014). Introducing depression and developmental screenings into the national programme on immunization (NPI) in southeast Nigeria: An experimental cross-sectional assessment. *General Hospital Psychiatry*, 36(1), 105–112. https://doi.org/10.1016/j.genhosppsych.2013.09.005

Beck, C. T., & Gable, R. K. (2001). Comparative analysis of the performance of the postpartum depression screening scale with two other depression instruments. *Nursing Research*, 50(4), 242–250. https://doi.org/10.1097/00006199-200107000-00008

Benedetti, A., Levis, B., Rücker, G., Jones, H. E., Schumacher, M., Ioannidis, J. P. A., Thombs, B., & DEPRESsion Screening Data (DEPRESSD) Collaboration. (2020). An empirical comparison of three methods for multiple cutoff diagnostic test meta-analysis of the Patient Health Questionnaire-9 (PHQ -9) depression screening tool using published data vs individual level data. *Research Synthesis Methods*, 11, 833. https://doi.org/10.1002/jrsm.1443

Bombardier, C. H., Kalpakjian, C. Z., Graves, D. E., Dyer, J. R., Tate, D. G., & Fann, J. R. (2012). Validity of the Patient Health Questionnaire-9 in assessing major depressive disorder during inpatient spinal cord injury rehabilitation. *Archives of Physical Medicine and Rehabilitation*, 93(10), 1838–1845. https://doi.org/10.1016/j.apmr.2012.04.019

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., Cohen, J. F., & STARD Group. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ (Clinical Research Ed)*, 351, h5527. https://doi.org/10.1136/bmj.h5527

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., & Standards for Reporting of Diagnostic Accuracy. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *BMJ (Clinical Research Ed)*, 326(7379), 41–44. https://doi.org/10.1136/bmj.326.7379.41

Bunevicius, A., Kusminskas, L., Pop, V. J., Pedersen, C. A., & Bunevicius, R. (2009). Screening for antenatal depression with the Edinburgh Depression Scale. *Journal of Psychosomatic Obstetrics & Gynecology*, 30(4), 238–243. https://doi.org/10.3109/01674820903230708

Cholera, R., Gaynes, B. N., Pence, B. W., Bassett, J., Qangule, N., Macphail, C., Brenhardt, S., Pettifor, A., & Miller, W. C. (2014). Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. *Journal of Affective Disorders*, 167, 160–166. https://doi.org/10.1016/j.jad.2014.06.003

Couto, T., Brancaglion, M. Y. M., Cardoso, M. N., Protzner, A. B., Garcia, F. D., Nicolato, R., Aguiar, R. A. L. P., Leite, H. V., & Corrêa, H. (2015). What is the best tool for screening antenatal depression? *Journal of Affective Disorders*, *178*, 12–17. https://doi.org/10.1016/j.jad.2015.02.003

Chaudron, L. H., Szilagyi, P. G., Tang, W., Anson, E., Talbot, N. L., Wadkins, H. I. M., Tu, X., & Wisner, K. L. (2010). Accuracy of depression screening tools for identifying postpartum depression among urban mothers. *Pediatrics*, *125*(3), e609–e617. https://doi.org/10.1542/peds.2008-3261

Cochrane methods: IPD meta-analysis. (2020). *Frequently asked questions.* Retrieved from https://methods.cochrane.org/ipdma/frequently-asked-questions#9

Cox, J. L., Holden, J. M., & Sagovsky, R. (1987). Detection of postnatal Depression. *British Journal of Psychiatry*, *150*(6), 782–786. https://doi.org/10.1192/bjp.150.6.782

Delgadillo, J., Payne, S., Gilbody, S., Godfrey, C., Gore, S., Jessop, D., & Dale, V. (2011). How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. *Journal of Affective Disorders*, *134*(1–3), 266–271. https://doi.org/10.1016/j.jad.2011.06.017

de Lima Osorio, F., Vilela Mendes, A., Crippa, J. A., & Loureiro, S. R. (2009). Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of Brazilian women in the context of primary health care. *Perspectives in Psychiatric Care*, *45*(3), 216–227. https://doi.org/10.1111/j.1744-6163.2009.00224.x

Fann, J. R., Bombardier, C. H., Dikmen, S., Esselman, P., Warms, C. A., Pelzer, E., Rau, H., & Temkin, N. (2005). Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *Journal of Head Trauma Rehabilitation*, *20*(6), 501–511. https://doi.org/10.1097/00001199-200511000-00003

Garcia-Esteve, L., Ascaso, C., Ojuel, J., & Navarro, P. (2003). Validation of the Edinburgh Postnatal Depression Scale (EPDS) in Spanish mothers. *Journal of Affective Disorders*, *75*(1), 71–76. https://doi.org/10.1016/s0165032702000204

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine*, *22*(11), 1596–1602. https://doi.org/10.1007/s11606-007-0333-y

Gjerdingen, D., Crow, S., McGovern, P., Miner, M., & Center, B. (2009). Postpartum depression screening at well-child visits: Validity of a 2-question screen and the PHQ-9. *Annals of Family Medicine*, *7*(1), 63–70. https://doi.org/10.1370/afm.933

Hewitt, C., Gilbody, S., Brealey, S., Paulden, M., Palmer, S., Mann, R., Green, J., Morrell, J., Barkham, M., Light, K., & Richards, D. (2009). Methods to identify postnatal depression in primary care: An integrated evidence synthesis and value of information analysis. *Health Technology Assessment*, *13*(36), 1–230. https://doi.org/10.3310/hta13360

Howard, L. M., Molyneaux, E., Dennis, C.-L., Rochat, T., Stein, A., & Milgrom, J. (2014). Non-psychotic mental disorders in the perinatal period. *The Lancet*, *384*(9956), 1775–1788. https://doi.org/10.1016/S0140-6736(14)61276-9

Hyphantis, T., Kotsis, K., Voulgari, P. V., Tsifetaki, N., Creed, F., & Drosos, A. A. (2011). Diagnostic accuracy, internal consistency, and convergent validity of the Greek version of the Patient Health Questionnaire 9 in diagnosing depression in rheumatologic disorders. *Arthritis Care & Research*, *63*(9), 1313–1321. https://doi.org/10.1002/acr.20505

Inagaki, M., Ohtsuki, T., Yonemoto, N., Kawashima, Y., Saitoh, A., Oikawa, Y., Kurosawa, M., Muramatsu, K., Furukawa, T. A., & Yamada, M. (2013). Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: A cross-sectional study. *General Hospital Psychiatry*, *35*(6), 592–597. https://doi.org/10.1016/j.genhosppsych.2013.08.001

Ioannidis, J. P. A., Rosenberg, P. S., Goedert, J. J., & O'Brien, T. R., & International Meta-analysis of HIV Host Genetics. (2002). Commentary: Meta-analysis of individual participants' data in genetic epidemiology. *American Journal of Epidemiology*, *156*(3), 204–210. https://doi.org/10.1093/aje/kwf031

Khalifa, D., Glavin, K., Bjertness, E., & Lien, L. (2015). Postnatal depression among Sudanese women: Prevalence and validation of the Edinburgh Postnatal Depression Scale at 3 months postpartum. *International Journal of Women's Health*, *7*, 677–684. https://doi.org/10.2147/IJWH.S81401

Khamseh, M. E., Baradaran, H. R., Javanbakht, A., Mirghorbani, M., Yadollahi, Z., & Malek, M. (2011). Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran. *BMC Psychiatry*, *11*, 61–244X. https://doi.org/10.1186/1471-244X-11-61

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, *340*, c365. https://doi.org/10.1136/bmj.c365

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lambert, S. D., Clover, K., Pallant, J. F., Britton, B., King, M. T., Mitchell, A. J., & Carter, G. (2015). Making sense of variations in prevalence estimates of depression in cancer: A co-calibration of commonly used depression scales using rasch analysis. *Journal of the National Comprehensive Cancer Network*, *13*(10), 1203–1211. https://doi.org/10.6004/jnccn.2015.0149

Leeflang, M. M. G., Moons, K. G. M., Reitsma, J. B., & Zwinderman, A. H. (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: Mechanisms, magnitude, and solutions. *Clinical Chemistry*, *54*(4), 729–737. https://doi.org/10.1373/clinchem.2007.096032

Levis, B., Benedetti, A., Levis, A. W., Ioannidis, J. P. A., Shrier, I., Cuijpers, P., Gilbody, S., Kloda, L. A., McMillan, D., Patten, S. B., Steele, R. J., Ziegelstein, R. C., Bombardier, C. H., de Lima Osório, F., Fann, J. R., Gjerdingen, D., Lamers, F., Lotrakul, M., Loureiro, S. R., … Thombs, B. D. (2017). Selective cutoff reporting in studies of diagnostic test accuracy: A comparison of conventional and individual-patient-data meta-analyses of the Patient Health Questionnaire-9 depression screening tool. *American Journal of Epidemiology*, *185*(10), 954–964. https://doi.org/10.1093/aje/kww191

Levis, B., Benedetti, A., & Thombs, B. D., & 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. https://doi.org/10.1136/bmj.l1476

Levis, B., Negeri, Z., Sun, Y., Benedetti, A., & Thombs, B. D., & DEPRESsion Screening Data (DEPRESSD) EPDS Group. (2020). Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: Systematic review and meta-analysis of individual participant data. *BMJ*, *371*, m4022. https://doi.org/10.1136/bmj.m4022

Lotrakul, M., Sumrithe, S., & Saipanish, R. (2008). Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry*, *8*, 46. https://doi.org/10.1186/1471-244X-8-46

Lowe, B., Spitzer, R. L., Grafe, K., Kroenke, K., Quenter, A., Zipfel, S., & Herzog, W. (2004). Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' Diagnoses. *Journal of Affective Disorders*, *78*(2), 131–140. https://doi.org/10.1016/s0165-0327(02)00237-9

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. *Canadian Medical Association Journal*, *184*(3), E191–E196. https://doi.org/10.1503/cmaj.110829

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of Clinical Epidemiology*, *75*, 40–46. https://doi.org/10.1016/j.jclinepi.2016.01.021

Moriarty, A. S., Gilbody, S., McMillan, D., & 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*(6), 567–576. https://doi.org/10.1016/j.genhosppsych.2015.06.012

O'Connor, E., Rossom, R. C., Henninger, M., Groom, H. C., & Burda, B. U. (2016). Primary care screening for and treatment of depression in pregnant and postpartum women: Evidence report and systematic review for the US preventive services task force. *JAMA*, *315*(4), 388–406. https://doi.org/10.1001/jama.2015.18948

Pawlby, S., Sharp, D., Hay, D., & O'Keane, V. (2008). Postnatal depression and child outcome at 11 years: The importance of accurate diagnosis. *Journal of Affective Disorders*, *107*(1–3), 241–245. https://doi.org/10.1016/j.jad.2007.08.002

Pence, B. W., Gaynes, B. N., Atashili, J., O'Donnell, J. K., Tayong, G., Kats, D., Whetten, R., Whetten, K., Njamnshi, A. K., & Ndumbe, P. M. (2012). Validity of an interviewer-administered Patient Health Questionnaire-9 to screen for depression in HIV-infected patients in Cameroon. *Journal of Affective Disorders*, *143*(1–3), 208–213. https://doi.org/10.1016/j.jad.2012.05.056

Phillips, J., Charles, M., Sharpe, L., & Matthey, S. (2009). Validation of the subscales of the Edinburgh Postnatal Depression Scale in a sample of women with unsettled infants. *Journal of Affective Disorders*, *118*(1–3), 101–112. https://doi.org/10.1016/j.jad.2009.02.004

Radoš, S. N., Tadinac, M., & Herman, R. (2013). Validation study of the Croatian version of the Edinburgh Postnatal Depression Scale (EPDS). *Suvremena Psihologija*, *16*(2), 203–208.

Riley, R. D., Dodd, S. R., Craig, J. V., Thompson, J. R., & Williamson, P. R. (2008). Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine*, *27*(29), 6111–6136. https://doi.org/10.1002/sim.3441

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, *340*, c221. https://doi.org/10.1136/bmj.c221

Rochat, T. J., Tomlinson, M., Newell, M.-L., & Stein, A. (2013). Detection of antenatal depression in rural HIV-affected populations with short and ultrashort versions of the Edinburgh Postnatal Depression Scale (EPDS). *Archives of Women's Mental Health*, *16*(5), 401–410. https://doi.org/10.1007/s00737-013-0353-z

Sherina, M. S., Arroll, B., & Goodyear-Smith, F. (2012). Criterion validity of the PHQ-9 (Malay version) in a primary care clinic in Malaysia. *The Medical Journal of Malaysia*, *67*(3), 309–315.

Spitzer, R. L., Kroenke, K., & Williams, J. B., & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MDThe PHQ primary care Study. *JAMA*, *282*(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Stafford, L., Berk, M., & Jackson, H. J. (2007). Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. *General Hospital Psychiatry*, *29*(5), 417–424. https://doi.org/10.1016/j.genhosppsych.2007.06.005

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions*, *25*(1), 76–97. https://doi.org/10.1177/0163278702025001006

Su, K.-P., Chiu, T.-H., Huang, C.-L., Ho, M., Lee, C.-C., Wu, P.-L., Lin, C.-Y., Liau, C.-H., Liao, C.-C., Chiu, W.-C., & Pariante, C. M. (2007). Different cutoff points for different trimesters? The use of Edinburgh Postnatal Depression Scale and Beck Depression Inventory to screen for depression in pregnant Taiwanese women. *General Hospital Psychiatry*, *29*(5), 436–441. https://doi.org/10.1016/j.genhosppsych.2007.05.005

Sung, S. C., Low, C. C. H., Fung, D. S. S., & Chan, Y. H. (2013). Screening for major and minor depression in a multiethnic sample of Asian primary care patients: A comparison of the nine-item Patient Health Questionnaire (PHQ-9) and the 16-item quick inventory of depressive symptomatology—Self-report (QIDS-SR16). *Asia-Pacific Psychiatry*, *5*(4), 249–258. https://doi.org/10.1111/appy.12101

Tandon, S. D., Cluxton-Keller, F., Leis, J., Le, H.-N., & Perry, D. F. (2012). A comparison of three screening tools to identify perinatal depression among low-income African American women. *Journal of Affective Disorders*, *136*(1–2), 155–162. https://doi.org/10.1016/j.jad.2011.07.014

Thiagayson, P., Krishnaswamy, G., Lim, M. L., Sung, S. C., Haley, C. L., Fung, D. S. S., Allen, J. C., & Chen, H. (2013). Depression and anxiety in Singaporean high-risk pregnancies—Prevalence and screening. *General Hospital Psychiatry*, *35*(2), 112–116. https://doi.org/10.1016/j.genhosppsych.2012.11.006

Thombs, B. D., Benedetti, A., Kloda, L. A., Levis, B., Nicolau, I., Cuijpers, P., Gilbody, S., Ioannidis, J. P. A., McMillan, D., Patten, S. B., Shrier, I., Steele, R. J., & Ziegelstein, R. C. (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*(1), 124. https://doi.org/10.1186/2046-4053-3-124

Thombs, B. D., Benedetti, A., Kloda, L. A., Levis, B., Riehm, K. E., Azar, M., Cuijpers, P., Gilbody, S., Ioannidis, J. P. A., McMillan, D., Patten, S. B., Shrier, I., Steele, R. J., Ziegelstein, R. C., Tonelli, M., Mitchell, N., Comeau, L., Schinazi, J., & Vigod, S. (2015). Diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for detecting major depression in pregnant and postnatal women: Protocol for a systematic review and individual patient data meta-analyses. *BMJ Open*, *5*(10), e009742. https://doi.org/10.1136/bmjopen-2015-009742

Thombs, B. D., Ziegelstein, R. C., & Whooley, M. A. (2008). Optimizing detection of major depression among patients with coronary artery disease using the patient health questionnaire: Data from the heart and soul study. *Journal of General Internal Medicine*, *23*(12), 2014–2017. https://doi.org/10.1007/s11606-008-0802-y

Toreki, A., Ando, B., Dudas, R. B., Dweik, D., Janka, Z., Kozinszky, Z., & Kereszturi, A. (2014). Validation of the Edinburgh Postnatal Depression Scale as a screening tool for postpartum depression in a clinical sample in Hungary. *Midwifery*, *30*(8), 911–918. https://doi.org/10.1016/j.midw.2014.02.008

Töreki, A., Andó, B., Keresztúri, A., Sikovanyecz, J., Dudas, R. B., Janka, Z., Kozinszky, Z., & Pál, A. (2013). The Edinburgh Postnatal Depression Scale: Translation and antepartum validation for a Hungarian sample. *Midwifery*, *29*(4), 308–315. https://doi.org/10.1016/j.midw.2012.01.011

Turner, A., Hambridge, J., White, J., Carter, G., Clover, K., Nelson, L., & Hackett, M. (2012). Depression screening in stroke: A comparison of alternative measures with the structured diagnostic interview for the diagnostic and statistical manual of mental disorders, fourth edition (major depressive episode) as criterion standard. *Stroke*, *43*(4), 1000–1005. https://doi.org/10.1161/STROKEAHA.111.643296

Twist, K., Stahl, D., Amiel, S. A., Thomas, S., Winkley, K., & Ismail, K. (2013). Comparison of depressive symptoms in type 2 diabetes using a two-stage survey design. *Psychosomatic Medicine*, *75*(8), 791–797. https://doi.org/10.1097/PSY.0b013e3182a2b108

Van der Leeden, R., Busing, F., & Meijer, E. (1997). *Bootstrap methods for two-level models. technical report PRM 97-04.* Department of Psychology, Leiden University.

Van der Leeden, R., Meijer, E., & Busing, F. (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 401–433). Springer. https://doi.org/10.1007/978-0-387-73186-5_11

van Steenbergen-Weijenburg, K. M., de Vroege, L., Ploeger, R. R., Brals, J. W., Vloedbeld, M. G., Veneman, T. F., … van der Feltz-Cornelis, C M. (2010). Validation of the PHQ-9 as a screening instrument for depression in diabetes patients in specialized outpatient clinics. *BMC Health Services Research*, 10, 235-6963. doi:10.1186/1472-6963-10-235

Vega-Dienstmaier, J. M., Mazzotti Suarez, G., & Campos Sanchez, M. (2002). Validation of a Spanish version of the Edinburgh Postnatal Depression Scale. *Actas Espanolas De Psiquiatria*, 30(2), 106–111.

Vohringer, P. A., Jimenez, M. I., Igor, M. A., Fores, G. A., Correa, M. O., Sullivan, M. C., Holtzman, N. S., Whitham, E. A., Barroilhet, S. A., Alvear, K., Logvinenko, T., Kent, D. M., & Ghaemi, N. S. (2013). Detecting mood disorder in resource-limited primary care settings: Comparison of a self-administered screening tool to general practitioner assessment. *Journal of Medical Screening*, 20(3), 118–124. https://doi.org/10.1177/0969141313503954

Wittkampf, K. A., Naeije, L., Schene, A. H., Huyser, J., & van Weert, H. C. (2007). Diagnostic accuracy of the mood module of the Patient Health Questionnaire: A systematic review. *General Hospital Psychiatry*, 29(5), 388–395. https://doi.org/10.1016/j.genhosppsych.2009.06.001

Wu, Y., Levis, B., Ioannidis, J. P. A., Benedetti, A., Thombs, B. D., & DEPRESsion Screening Data (DEPRESSD) Collaboration. (2021). Probability of major depression classification based on the SCID, CIDI, and MINI diagnostic interviews: A synthesis of three individual participant data meta-analyses. *Psychotherapy and Psychosomatics*, 90(1), 28–40. https://doi.org/10.1159/000509283

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.