Psychotherapy and Psychosomatics

The Accuracy of the Patient Health Questionnaire-9 Algorithm for Screening to Detect Major Depression: An Individual Participant Data Meta-Analysis

Chen He[a, b]   Brooke Levis[a, b]   Kira E. Riehm[a]   Nazanin Saadat[a]   Alexander W. Levis[a, b]   Marleine Azar[a, b]
Danielle B. Rice[a, c]   Ankur Krishnan[a]   Yin Wu[a, b, d]   Ying Sun[a]   Mahrukh Imran[a]   Jill Boruff[e]   Pim Cuijpers[f]
Simon Gilbody[g]   John P.A. Ioannidis[h]   Lorie A. Kloda[i]   Dean McMillan[g]   Scott B. Patten[j, k]   Ian Shrier[a, b]
Roy C. Ziegelstein[l]   Dickens H. Akena[m]   Bruce Arroll[n]   Liat Ayalon[o]   Hamid R. Baradaran[p, r]   Murray Baron[a, q]
Anna Beraldi[s]   Charles H. Bombardier[t]   Peter Butterworth[u, v]   Gregory Carter[w]   Marcos Hortes Nisihara Chagas[y]
Juliana C.N. Chan[y, z, A]   Rushina Cholera[B]   Kerrie Clover[C, D]   Yeates Conwell[E]   Janneke M. de Man-van Ginkel[F]
Jesse R. Fann[G]   Felix H. Fischer[H]   Daniel Fung[I, J, K, L]   Bizu Gelaye[M]   Felicity Goodyear-Smith[n]
Catherine G. Greeno[N]   Brian J. Hall[O, P]   Patricia A. Harrison[Q]   Martin Härter[R]   Ulrich Hegerl[S]   Leanne Hides[T]
Stevan E. Hobfoll[U]   Marie Hudson[a, q]   Thomas N. Hyphantis[V]   Masatoshi Inagaki[W]   Khalida Ismail[X]
Nathalie Jetté[j, k, Y]   Mohammad E. Khamseh[P]   Kim M. Kiely[Z, α]   Yunxin Kwan[β]   Femke Lamers[Y]   Shen-Ing Liu[L, δ, ε, ζ]
Manote Lotrakul[n]   Sonia R. Loureiro[Y]   Bernd Löwe[I]   Laura Marsh[k]   Anthony McGuire[λ]   Sherina Mohd-Sidik[µ]
Tiago N. Munhoz[v]   Kumiko Muramatsu[ξ]   Flávia L. Osório[Y, o]   Vikram Patel[π, ρ]   Brian W. Pence[σ]
Philippe Persoons[τ, υ]   Angelo Picardi[φ]   Katrin Reuter[X]   Alasdair G. Rooney[Ψ]   Iná S. da Silva dos Santos[v]
Juwita Shaaban[ω]   Abbey Sidebottom[ä]   Adam Simning[E]   Lesley Stafford[ö, ü]   Sharon Sung[I, L]
Pei Lin Lynnette Tan[β]   Alyna Turner[Ä, Ö]   Henk C.P.M. van Weert[Ü]   Jennifer White[+]   Mary A. Whooley[*, #, ±]
Kirsty Winkley[§]   Mitsuhiko Yamada[!]   Brett D. Thombs[a, b, c, d, q, $]   Andrea Benedetti[b, q, ‡]

[a]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, QC, Canada; [b]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada; [c]Department of Psychology, McGill University, Montréal, QC, Canada; [d]Department of Psychiatry, McGill University, Montréal, QC, Canada; [e]Schulich Library of Physical Sciences, Life Sciences and Engineering, McGill University, Montreal, QC, Canada; [f]Department of Clinical, Neuro- and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit, Amsterdam, The Netherlands; [g]Hull York Medical School and the Department of Health Sciences, University of York, Heslington, UK; [h]Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, CA, USA; [i]Library, Concordia University, Montréal, QC, Canada; [j]Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada; [k]Hotchkiss Brain Institute and O'Brien Institute for Public Health, University of Calgary, Calgary, AB, Canada; [l]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [m]Department of Psychiatry, Makerere University College of Health Sciences, Kampala, Uganda; [n]Department of General Practice and Primary Health Care, University of Auckland, Auckland, New Zealand; [o]Louis and Gabi Weisfeld School of Social Work, Bar Ilan University, Ramat Gan, Israel; [p]Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; [q]Department of Medicine, McGill University, Montréal, QC, Canada; [r]Ageing Clinical and Experimental Research Team, Institute of Applied Health Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, UK; [s]Kbo-Lech-Mangfall-Klinik Garmisch-Partenkirchen, Klinik für Psychiatrie, Psychotherapie und Psychosomatik, Lehrkrankenhaus der Technischen Universität München, Munich, Germany; [t]Department of Rehabilitation Medicine, University of Washington, Seattle, WA, USA; [u]Centre for Research on Ageing,

Brett D. Thombs, PhD
Jewish General Hospital
4333 Cote Ste Catherine Road
Montreal QC H3T 1E4 (Canada)
E-Mail brett.thombs@mcgill.ca

Andrea Benedetti, PhD
Centre for Outcomes Research and Evaluation, Research Institute of the McGill University Health Centre
5252 Boulevard de Maisonneuve
Montréal, QC H4A 3S5 (Canada)
E-Mail andrea.benedetti@mcgill.ca

Health and Wellbeing, Research School of Population Health, The Australian National University, Canberra, ACT, Australia; [v]Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Melbourne, VIC, Australia; [w]Centre for Brain and Mental Health Research, University of Newcastle, Newcastle, NSW, Australia; [x]Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; [y]Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR, China; [z]Asia Diabetes Foundation, Prince of Wales Hospital, Hong Kong SAR, China; [A]Hong Kong Institute of Diabetes and Obesity, Hong Kong SAR, China; [B]Department of Pediatrics, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC, USA; [C]Centre for Brain and Mental Health Research, University of Newcastle, Newcastle, NSW, Australia; [D]Psycho-Oncology Service, Calvary Mater Newcastle, Newcastle, NSW, Australia; [E]Department of Psychiatry, University of Rochester Medical Center, Rochester, NY, USA; [F]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; [G]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA; [H]Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, Berlin, Germany; [I]Department of Child and Adolescent Psychiatry, Institute of Mental Health, Singapore, Singapore; [J]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore; [K]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore; [L]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore; [M]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [N]School of Social Work, University of Pittsburgh, Pittsburgh, PA, USA; [O]Global and Community Mental Health Research Group, Department of Psychology, Faculty of Social Sciences, University of Macau, Macau SAR, China; [P]Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; [Q]City of Minneapolis Health Department, Minneapolis, MN, USA; [R]Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [S]Department of Psychiatry, Psychosomatics and Psychotherapy, Goethe-Universität Frankfurt, German Depression Foundation, Frankfurt, Germany; [T]School of Psychology, University of Queensland, Brisbane, QLD, Australia; [U]STAR-Stress, Anxiety and Resilience Consultants, Chicago, IL, USA; [V]Faculty of Medicine, School of Health Sciences, University of Ioannina, Ioannina, Greece; [W]Department of Psychiatry, Faculty of Medicine, Shimane University, Shimane, Japan; [X]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neurosciences, King's College London Weston Education Centre, London, UK; [Y]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [Z]School of Psychology, University of New South Wales, Sydney, NSW, Australia; [α]Neuroscience Research Australia, Sydney, NSW, Australia; [β]Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore, Singapore; [γ]Department of Psychiatry, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands; [δ]Department of Psychiatry, Mackay Memorial Hospital, Taipei, Taiwan; [ε]Department of Medical Research, Mackay Memorial Hospital, Taipei, Taiwan; [ζ]Department of Medicine, Mackay Medical College, Taipei, Taiwan; [η]Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; [ι]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [κ]Baylor College of Medicine Houston and Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX, USA; [λ]Department of Nursing, St. Joseph's College, Standish, ME, USA; [μ]Cancer Resource and Education Centre, and Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Malaysia; [ν]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil; [ξ]Department of Clinical Psychology, Graduate School of Niigata Seiryo University, Niigata, Japan; [o]National Institute of Science and Technology, Translational Medicine, Ribeirão Preto, Brazil; [π]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, USA; [ρ]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [σ]Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [τ]Department of Adult Psychiatry, University Hospitals Leuven, Leuven, Belgium; [υ]Department of Neurosciences, Katholieke Universiteit Leuven, Leuven, Belgium; [φ]Centre for Behavioural Sciences and Mental Health, Italian National Institute of Health, Rome, Italy; [χ]Practice for Psychotherapy and Psycho-Oncology, Freiburg, Germany; [ψ]Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburg, Edinburgh, UK; [ω]Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia; [ä]Allina Health, Minneapolis, MN, USA; [ö]Centre for Women's Mental Health, Royal Women's Hospital, Parkville, VIC, Australia; [ü]Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, VIC, Australia; [Ä]School of Medicine and Public Health, University of Newcastle, Newcastle, NSW, Australia; [Ö]IMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, VIC, Australia; [Ü]Department of General Practice, Amsterdam Institute for General Practice and Public Health, Amsterdam University Medical Centers, Amsterdam, The Netherlands; [+]Monash University, Melbourne, VIC, Australia; [*]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA; [#]Department of Medicine, Veterans Affairs Medical Center, San Francisco, CA, USA; [±]Department of Medicine, University of California San Francisco, San Francisco, CA, USA; [§]Florence Nightingale Faculty of Nursing, Midwifery and Palliative Care, King's College London, London, UK; [ı]Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Kodaira, Japan; [$]Department of Educational and Counselling Psychology, McGill University, Montréal, QC, Canada; [‡]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, QC, Canada

He et al.

Abstract

Background: Screening for major depression with the Patient Health Questionnaire-9 (PHQ-9) can be done using a cutoff or the PHQ-9 diagnostic algorithm. Many primary studies publish results for only one approach, and previous meta-analyses of the algorithm approach included only a subset of primary studies that collected data and could have published results. *Objective:* To use an individual participant data meta-analysis to evaluate the accuracy of two PHQ-9 diagnostic algorithms for detecting major depression and compare accuracy between the algorithms and the standard PHQ-9 cutoff score of ≥10. *Methods:* Medline, Medline In-Process and Other Non-Indexed Citations, PsycINFO, Web of Science (January 1, 2000, to February 7, 2015). Eligible studies that classified current major depression status using a validated diagnostic interview. *Results:* Data were included for 54 of 72 identified eligible studies (*n* participants = 16,688, *n* cases = 2,091). Among studies that used a semi-structured interview, pooled sensitivity and specificity (95% confidence interval) were 0.57 (0.49, 0.64) and 0.95 (0.94, 0.97) for the original algorithm and 0.61 (0.54, 0.68) and 0.95 (0.93, 0.96) for a modified algorithm. Algorithm sensitivity was 0.22–0.24 lower compared to fully structured interviews and 0.06–0.07 lower compared to the Mini International Neuropsychiatric Interview. Specificity was similar across reference standards. For PHQ-9 cutoff of ≥10 compared to semi-structured interviews, sensitivity and specificity (95% confidence interval) were 0.88 (0.82–0.92) and 0.86 (0.82–0.88). *Conclusions:* The cutoff score approach appears to be a better option than a PHQ-9 algorithm for detecting major depression.

© 2019 S. Karger AG, Basel

Introduction

Tests may be used for many purposes, including, for example, to discriminate between people who have improved versus not improved with treatment or to determine whether people suspected of having a condition may meet diagnostic criteria. Screening, however, is specifically done to attempt to identify a condition among apparently healthy people who are not suspected of having the condition [1, 2]. In depression screening, self-report symptom questionnaires are used to identify patients who have not been previously recognized as having a mental health problem, but who may have depression. Consistent with a clinimetric approach [3–5], in screening, patients who score above a pre-established cutoff threshold would need to be evaluated by a trained clinician to determine whether they have major depression and, if appropriate, offered treatment [6–10]. This assessment would include considerations that go beyond information obtained from the symptom questionnaire and would include consideration of the full set of diagnostic criteria, as well as contextual information, including function in daily life and performance of social roles and stressors, for instance [3–5]. Clinimetric approaches focus on sensitivity and specificity in relation to discriminating between different groups of patients and in terms of sensitivity to detecting changes in clinical or experimental settings; studies of screening test accuracy focus on discrimination between patients with and without a condition [3–5].

The Patient Health Questionnaire-9 (PHQ-9) [11–13], a 9-item self-report questionnaire, is the most commonly used depression screening tool in primary care [14]. Its 9 items align with the 9 Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria for a major depressive episode [15–17]. Item response options for each item range from "not at all" (score of 0) to "nearly every day" (score of 3), reflecting how often each symptom has bothered the respondent over the past 2 weeks. The PHQ-9 has been recommended by the United States Preventive Services Task Force (USPSTF) and others for depression screening in primary care and other settings, but recommendations do not specify the scoring approach to use [8, 18, 19]. Common approaches for screening include (1) a score cutoff threshold of ≥10 and (2) a diagnostic algorithm, which requires 5 or more items with scores of ≥2 points, with at least one being depressed mood or anhedonia [12]. Some researchers have used a modified algorithm that requires only 1 point for item 9 ("thoughts that you would be better off dead or of hurting yourself in some way") [13].

We have recently conducted an individual participant data meta-analysis (IPDMA) of PHQ-9 accuracy using the cutoff threshold approach (*n* studies = 58; *n* participants = 17,357) [20]. Compared to diagnoses made by semi-structured interviews, sensitivity and specificity for the standard cutoff of ≥10 (95% CI) for major depression were 0.88 (0.83–0.92) and 0.85 (0.82–0.88), respectively. A 2015 conventional meta-analysis of the diagnostic algorithm found that pooled sensitivity and specificity were 0.58 (0.50–0.66) and 0.94 (0.92–0.96) [21]. However, that study was based on only 27 primary studies and did not include results from 20 other studies that published re-

sults for the cutoff but not the algorithm [21, 22]. Other limitations were that it: (1) pooled results without distinguishing between the original PHQ-9 diagnostic algorithm, a modified algorithm, and other less frequently used algorithms; (2) could not evaluate accuracy in participant subgroups other than care setting, since primary studies did not report subgroup results; (3) could not exclude participants already diagnosed with depression who would not be screened in practice, but who were included in many primary studies [23, 24]; and (4) combined results across different types of reference standards, despite their inherent differences [20, 25].

IPDMA, which involves synthesis of participant-level data, rather than published summary results [26], allows the calculation of both cutoff and algorithm results and the conduct of subgroup analyses, even if not reported in the original studies. The objectives of the present IPDMA were to evaluate the diagnostic accuracy of the original and a modified PHQ-9 diagnostic algorithm: (1) among studies using different types of diagnostic interviews as reference standards, separately; (2) comparing participants not currently diagnosed or receiving treatment for a mental health problem to all patients regardless of diagnostic or treatment status; and (3) among subgroups based on age, sex, country human development index, and recruitment setting. We also compared accuracy results from the algorithms to results using the standard cutoff of ≥10.

Materials and Methods

This IPDMA was registered in PROSPERO (CRD42014010673), and a protocol was published [27]. Results were reported per PRISMA-DTA [28] and PRISMA-IPD [29] statements.

Study Eligibility

Data sets from articles in any language were eligible if they included diagnostic classification for current major depressive disorder (MDD) or major depressive episode (MDE) based on a validated semi-structured or fully structured interview conducted within 2 weeks of PHQ-9 administration among participants ≥18 years not recruited from youth or psychiatric settings or pre-identified as having depressive symptoms. We required the interviews and PHQ-9 to be administered within 2 weeks of each other to be consistent with DSM [15–17] and International Classification of Diseases (ICD) [30] major depression diagnostic criteria. We excluded patients from psychiatric settings or already identified as having depressive symptoms because screening is done to identify unrecognized cases.

Data sets where not all participants were eligible were included if primary data allowed selection of eligible participants. For defining major depression, we considered MDD or MDE based on the DSM or ICD. If more than one was reported, we prioritized DSM

over ICD and MDE over MDD, because screening would detect episodes and then determine whether the episode is related to MDD or bipolar disorder based on further assessment. Across all studies, there were only 23 discordant diagnoses depending on classification prioritization (0.1% of participants). For the present study, in order to be able to evaluate accuracy of both PHQ-9 diagnostic algorithms, we only included primary studies with databases that provided individual PHQ-9 item scores and not just PHQ-9 total scores.

Database Searches and Study Selection

A medical librarian searched Medline, Medline In-Process and Other Non-Indexed Citations via Ovid, PsycINFO, and Web of Science (January 1, 2000, to February 7, 2015), using a peer-reviewed [31] search strategy (online suppl. Methods 1; for all online suppl. material, see www.karger.com/doi/10.1159/000502294) limited to the year 2000 forward because the PHQ-9 was published in 2001 [11]. We also reviewed reference lists of relevant reviews and queried contributing authors about nonpublished studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, ON, Canada).

Two investigators independently reviewed titles and abstracts for eligibility. 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 in which team members were fluent.

Data Extraction and Synthesis

Authors of eligible data sets were invited to contribute de-identified primary data. We emailed corresponding authors of eligible primary studies at least 3 times, as necessary. If we did not receive a response, we emailed co-authors and attempted to contact corresponding authors by phone.

Country, recruitment setting (nonmedical, primary care, inpatient, outpatient specialty), and diagnostic interview were extracted from published reports by two investigators independently, with disagreements resolved by consensus. Countries were categorized as "very high" or "other" development based on the United Nation's human development index, a statistical composite index that includes indicators of life expectancy, education, and income [32]. Participant-level data included age, sex, major depression status, current mental health diagnosis or treatment, and PHQ-9 item scores. In two primary studies, multiple recruitment settings were included; thus, recruitment setting was coded at the participant level. When data sets included statistical weights to reflect sampling procedures, we used the weights provided. For studies where sampling procedures merited weighting, but the original study did not, we constructed weights using inverse selection probabilities. Weighting occurred, for instance, when all participants with positive screens and a random subset of participants with negative screens were administered a diagnostic interview.

Two investigators assessed risk of bias of included studies independently, based on primary publications, using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool. Discrepancies were resolved by consensus. See supplementary Methods 2 for coding rules [33].

Individual participant data were converted to a standard format and synthesized into a single data set with study-level data. We

compared published participant characteristics and diagnostic accuracy results with results from raw data sets and resolved any discrepancies in consultation with study investigators.

Data Analysis

We conducted three sets of analyses. First, we estimated sensitivity and specificity for the original and modified PHQ-9 diagnostic algorithms for all patients, separately by studies that used semi-structured (Structured Clinical Interview for DSM [SCID] [34], Schedules for Clinical Assessment in Neuropsychiatry [SCAN] [35], Depression Interview and Structured Hamilton [DISH] [36]), fully structured (Composite International Diagnostic Interview [CIDI] [37], Clinical Interview Schedule-Revised [CIS-R] [38], Diagnostic Interview Schedule [DIS] [39]), and Mini International Neuropsychiatric Interview (MINI) [40, 41] reference standards. This is because in a recent analysis, we found that the MINI classified approximately twice as many participants with major depression as the CIDI controlling for depressive symptom scores [25]. Compared to semi-structured interviews, fully structured interviews (MINI excluded) classified more patients with low symptom scores but fewer patients with high symptom scores. These findings are consistent with the design of each type of reference standard. Semi-structured diagnostic interviews are intended for administration by experienced diagnosticians, require clinical judgment, and allow rephrasing of questions and follow-up probes. Fully structured interviews are designed to be administered by lay interviewers, are fully scripted, and do not allow deviation. They are intended to achieve a high level of standardization but may sacrifice accuracy [42–45]. The MINI is fully structured but was designed for very rapid administration and was described by its authors as intended to be overinclusive [40, 41].

Second, for each reference standard category, we estimated sensitivity and specificity for the original and modified diagnostic algorithms, only including participants not currently diagnosed or receiving mental health treatment, and we compared results to those for all participants. This was done because existing conventional meta-analyses have been based on primary studies that typically do not exclude patients already diagnosed or receiving treatment, but who would not be screened in practice, since screening is done to identify unrecognized cases [23, 24].

Third, for each reference standard category, we compared sensitivity and specificity of the original and modified diagnostic algorithms among subgroups based on age (<60 vs. ≥60 years), sex, country human development index, and recruitment setting. For the MINI, we combined inpatient and outpatient specialty care settings, because only one study included inpatients. We excluded primary studies with no major depression cases in subgroup analyses since this did not allow application of the bivariate random effects model. A maximum of 15 participants were excluded from any subgroup analysis.

For each meta-analysis, bivariate random-effects models were fitted via Gauss-Hermite quadrature [46]. This 2-stage meta-analytic approach modeled sensitivity and specificity simultaneously, accounting for the inherent correlation between them and for precision of estimates within studies. For each analysis, this model provided estimates of pooled sensitivity and specificity.

We estimated differences in sensitivity and specificity between subgroups for the original and modified diagnostic algorithms by constructing confidence intervals (CIs) for differences

via a clustered bootstrap approach [47, 48], resampling at the study and participant levels. For each comparison, we ran 1,000 iterations.

For heterogeneity estimation, we generated forest plots of sensitivities and specificities for the original and modified diagnostic algorithms for each study, first for all studies in each reference standard category, and then separately across participant subgroups within each reference standard category. In addition, we quantified heterogeneity overall and for subgroups by reporting estimated variances of random effects for sensitivity and specificity (τ^2) and by estimating R, the ratio of the estimated standard deviation of the pooled sensitivity (or specificity) from the random-effects model to that from the corresponding fixed-effects model [49]. We used a complete case analysis since only 2% of participants were missing PHQ-9 item data or covariate data.

To estimate positive and negative predictive values using the original and modified algorithms, we generated nomograms and applied sensitivity and specificity estimates from the meta-analysis to hypothetical major depression prevalence values of 5–25%.

In sensitivity analyses, for each reference standard category, we compared accuracy results across subgroups based on QUADAS-2 items with at least 100 major depression cases among participants in studies categorized as having "low" risk of bias and in studies with "high" or "unclear" risk of bias.

We previously published an IPDMA of the accuracy of the PHQ-9 using the cutoff threshold approach for screening to detect major depression (*n* studies = 58; *n* participants = 17,357) and found that accuracy was highest compared to diagnoses made by semi-structured interviews; sensitivity and specificity for the standard cutoff of ≥10 (95% CI) were 0.88 (0.83–0.92) and 0.85 (0.82–0.88) [20]. That IPDMA included data from 4 primary studies that could not be included in the present IPDMA because the individual PHQ-9 item scores needed to apply the diagnostic algorithms were not available. To ensure that we could directly compare results from the PHQ-9 diagnostic algorithm to published results using the cutoff threshold approach, we re-evaluated sensitivity and specificity for the cutoff approach with the standard cutoff of ≥10 using the same data set used for the present evaluation of the diagnostic algorithms (*n* = 16,688).

All analyses were run in R (R version 3.4.1 and R Studio version 1.0.143) using the glmer function within the lme4 package, which uses one quadrature point.

Results

Search Results and Inclusion of Primary Data

Of 5,248 unique titles and abstracts identified from the database search, 5,039 were excluded after title and abstract review and 113 after full-text review, leaving 96 eligible articles with data from 69 unique participant samples, of which 55 (80%) contributed data sets (online suppl. Fig. 1). In addition, authors of included studies contributed data from 3 unpublished studies (2 subsequently published) [50, 51], for a total of 58 data sets of 72 identified eligible data sets (81%). Of these, 4 studies

Table 1. Participant data by diagnostic interview

Diagnostic interview	Studies	Participants	Major depression	
			n	%
Semi-structured				
SCID	24	4,347	649	15
SCAN	2	1,884	129	7
DISH	1	100	9	9
Fully structured				
CIDI	11	6,272	554	9
DIS	1	1,006	221	22
CIS-R	1	299	13	4
MINI	14	2,780	516	19
Total	54	16,688	2,091	13

CIDI, Composite International Diagnostic Interview; CIS-R, Clinical Interview Schedule-Revised; DIS, Diagnostic Interview Schedule; DISH, Depression Interview and Structured Hamilton; MINI, Mini International Neuropsychiatric Interview; SCAN, Schedules for Clinical Assessment in Neuropsychiatry; SCID, Structured Clinical Interview for DSM Disorders.

contributed PHQ-9 total scores but did not provide item-level data and were excluded; thus, there were 54 studies with 17,050 participants. From those 54 studies, we excluded 308 participants who were missing PHQ-9 item scores and 54 participants who were missing covariate data, leaving 16,688 participants (2,091 major depression cases [13%]) who were included in analyses (77% of eligible participants). Reasons for exclusion for all articles excluded at full-text level and characteristics of included studies and those that did not provide data for the present study are shown in online supplementary Table 1, online supplementary Table 2a, and online supplementary Table 2b.

Of the 54 included studies, 27 used semi-structured reference standards, 13 used fully structured reference standards (MINI excluded), and 14 used the MINI (Table 1). The SCID was the most common semi-structured interview (24 studies, 4,347 participants), and the CIDI was the most common fully structured interview (11 studies, 6,272 participants). Among studies that used semi-structured, fully structured, and MINI diagnostic interviews, mean sample sizes were 234, 583, and 199, and mean numbers (%) with major depression were 29 (12%), 61 (10%), and 37 (19%). Characteristics of participants are shown in Table 2.

PHQ-9 Diagnostic Algorithm Accuracy by Reference Standard

Comparisons of sensitivity and specificity estimates by reference standard category are shown in Table 3. Sensitivity and specificity estimates for the original and modified algorithms differed by 0.04 or less within each reference standard category. Compared to semi-structured interviews, sensitivity and specificity were 0.57 (95% CI, 0.49–0.64) and 0.95 (95% CI, 0.94–0.97) for the original algorithm and 0.61 (95% CI, 0.54–0.68) and 0.95 (0.93–0.96) for the modified algorithm. Specificity was similar for studies that compared PHQ-9 algorithms to semi-structured interviews, fully structured interviews, or the MINI. Sensitivity, however, was substantially higher compared to semi-structured interviews than compared to fully structured interviews or the MINI (Table 4). Heterogeneity analyses suggested moderate heterogeneity across studies. For original and modified diagnostic algorithms, sensitivity and specificity forest plots are shown in online supplementary Figures 2a–f and 3a–f, with τ^2 and R values shown in online supplementary Table 3.

Nomograms of positive and negative predictive values for the original and modified algorithms for hypothetical major depression prevalence values of 5–25% are shown in online supplementary Figures 4a, b and 5a, b. For the prevalence of studies included in the IPDMA (13%), for the original diagnostic algorithm, positive and negative predictive values for semi-structured, fully structured (MINI excluded) and MINI were 0.63 and 0.94, 0.51 and 0.91, and 0.72 and 0.93, respectively; for the modified algorithm, they were 0.65 and 0.94, 0.53 and 0.91, and 0.67 and 0.93, respectively.

PHQ-9 Diagnostic Algorithm Accuracy among Participants Not Diagnosed or Receiving Treatment for a Mental Health Problem Compared to All Participants

Sensitivity and specificity estimates were not statistically significantly different for any reference standard category when restricted to participants not currently diagnosed or receiving treatment for a mental health problem compared to all participants. See online supplementary Table 4 for results.

PHQ-9 Diagnostic Algorithm Accuracy among Subgroups

For each reference standard category, comparisons of sensitivity and specificity estimates of the original PHQ-9 diagnostic algorithm and the modified PHQ-9 diagnostic algorithm among all participants and among participant

Table 2. Participant data by subgroup

Participant subgroup	Semi-structured diagnostic interviews			Fully structured diagnostic interviews			Mini International Neuropsychiatric Interview		
	studies	participants	major depression, n (%)	studies	participants	major depression, n (%)	studies	participants	major depression, n (%)
All participants	27	6,331	787 (12)	13	7,577	788 (10)	14	2,780	516 (19)
Participants not currently diagnosed or receiving treatment for a mental health problem	18	2,596	305 (12)	5	4,085	272 (7)	6	927	168 (18)
Age <60 years	24	3,844	516 (13)	13	5,401	594 (11)	13	1,885	290 (15)
Age ≥60 years	24	2,480	271 (11)	10	2,175	194 (9)	12	880	226 (26)
Women	26	3,671	477 (13)	13	4,261	456 (11)	14	1,580	325 (21)
Men	23	2,653	310 (12)	12	3,316	332 (10)	14	1,200	191 (16)
Very high country human development index	23	5,863	634 (11)	8	5,637	541 (10)	9	1,752	397 (23)
Low, medium or high country human development index	4	468	153 (33)	5	1,940	247 (13)	5	1,028	119 (12)
Nonmedical care	2	567	105 (19)	2	963	74 (8)	2	299	72 (24)
Primary care	8	2,965	317 (11)	5	3,578	273 (8)	5	1,290	168 (13)
Inpatient specialty care	7	733	76 (10)	2	372	34 (9)	1	137	25 (18)
Outpatient specialty care	12	2,066	289 (14)	4	2,664	407 (15)	7	1,054	251 (24)

Some variables were coded at the study level, while others were coded at the participant level. Thus, number of studies does not always add up to the total for each reference standard.

Table 3. Comparison of sensitivity and specificity estimates for original[a] and modified[b] PHQ-9 diagnostic algorithms for each reference standard category

Algorithm	Semi-structured reference standard[c]				Fully structured reference standard[d]				MINI reference standard[e]			
	sensitivity	95% CI	specificity	95% CI	sensitivity	95% CI	specificity	95% CI	sensitivity	95% CI	specificity	95% CI
Original PHQ-9 diagnostic algorithm	0.57	0.49, 0.64	0.95	0.94, 0.97	0.35	0.26, 0.46	0.95	0.93, 0.97	0.51	0.49, 0.53	0.97	0.96, 0.98
Modified PHQ-9 diagnostic algorithm	0.61	0.54, 0.68	0.95	0.93, 0.96	0.37	0.27, 0.48	0.95	0.92, 0.97	0.54	0.45, 0.64	0.96	0.94, 0.97
Cutoff of ≥10	0.88	0.82, 0.92	0.86	0.82, 0.88	0.67	0.57, 0.76	0.86	0.80, 0.90	0.75	0.66, 0.82	0.88	0.84, 0.91

PHQ-9, Patient Health Questionnaire-9; CI, confidence interval; MINI, Mini International Neuropsychiatric Interview. [a] The original diagnostic algorithm requires 5 or more items with scores of ≥2 points, with at least 1 being depressed mood or anhedonia. [b] The modified algorithm requires 5 or more items with scores of ≥2 points (except item 9 that only requires 1 point), with at least 1 being depressed mood or anhedonia. [c] Studies = 27; participants = 6,331; major depression = 787. [d] Studies = 13; participants = 7,577; major depression = 788. [e] Studies = 14; participants = 2,780; major depression = 516.

Table 4. Differences in sensitivity and specificity estimates for original[a] and modified[b] PHQ-9 diagnostic algorithms for fully structured reference standards and MINI reference standards compared to semi-structured reference standards

Algorithm	Semi-structured minus fully structured[c, d]				Semi-structured minus MINI[e, f]			
	sensitivity	95% CI	specificity	95% CI	sensitivity	95% CI	specificity	95% CI
Original PHQ-9 diagnostic algorithm	0.22	–0.04, 0.36	0.00	–0.02, 0.02	0.06	–0.04, 0.16	–0.02	–0.03, 0.02
Modified PHQ-9 diagnostic algorithm	0.24	–0.03, 0.36	0.00	–0.04, 0.04	0.07	–0.03, 0.16	–0.02	–0.03, 0.02

PHQ-9, Patient Health Questionnaire-9; CI, confidence interval; MINI, Mini International Neuropsychiatric Interview. [a] The original diagnostic algorithm requires 5 or more items with scores of ≥2 points, with at least 1 being depressed mood or anhedonia. [b] The modified algorithm requires 5 or more items with scores of ≥2 points (except item 9 that only requires 1 point), with at least 1 being depressed mood or anhedonia. [c] One bootstrap iteration (0.01%) did not produce a difference estimate for the PHQ-9 diagnostic algorithm. This iteration was removed prior to determining the bootstrapped CI. [d] One bootstrap iteration (0.01%) did not produce a difference estimate for the modified PHQ-9 diagnostic algorithm. This iteration was removed prior to determining the bootstrapped CI. [e] One bootstrap iteration (0.01%) did not produce a difference estimate for the PHQ-9 diagnostic algorithm. This iteration was removed prior to determining the bootstrapped CI. [f] One bootstrap iteration (0.01%) did not produce a difference estimate for the modified PHQ-9 diagnostic algorithm. This iteration was removed prior to determining the bootstrapped CI.

subgroups based on age, sex, human development index, and care setting are shown in online supplementary Table 4, with forest plots shown in online supplementary Figures 2a–f and 3a–f, and τ^2 and R values shown in online supplementary Table 3. Overall, there were no examples of statistically significant differences in diagnostic accuracy across subgroups that were replicated in more than a single reference standard category. Heterogeneity improved in some instances when subgroups were considered.

Risk of Bias Sensitivity Analyses

Online supplementary Table 5 shows QUADAS-2 ratings for each included primary study. There were no significant or substantive differences based on QUADAS-2 ratings that were replicated across reference standards.

Sensitivity and Specificity of PHQ-9 Using a Cutoff Threshold of ≥10

Based on the same data set as used with the diagnostic algorithm analyses ($n = 16,688$), compared to a semi-structured diagnostic interview, sensitivity and specificity for a cutoff of ≥10 (95% CI) were 0.88 (0.82–0.92) and 0.86 (0.82–0.88). For fully structured interviews (MINI excluded), sensitivity and specificity were 0.67 (0.57–0.76) and 0.86 (0.80–0.90). For the MINI, sensitivity and specificity were 0.75 (0.66–0.82) and 0.88 (0.84–0.91).

Discussion

Conventional meta-analyses on the accuracy of the PHQ-9 for screening that have used either the cutoff threshold or diagnostic algorithm approaches have been limited because most primary studies publish results from one, but not both, approaches. By using IPDMA, we were able to analyze data from twice as many primary studies as were included in the most recent meta-analysis of the PHQ-9 diagnostic algorithm (54 vs. 27) [21] and to directly compare results to a cutoff score of ≥10 using the same data.

The main finding was that for both the original and modified PHQ-9 diagnostic algorithms, sensitivity was low across reference standards and subgroups, although specificity was high. Sensitivity and specificity to distinguish between people with and without a condition is a core clinimetric requirement [3–5]. Sensitivity was 0.57 (specificity = 0.95) for the original algorithm and 0.61 (specificity = 0.95) for the modified algorithm compared to semi-structured diagnostic interviews; accuracy was poorer compared to fully structured interviews or the MINI. We found no differences in accuracy by subgroups that were consistent across reference standards. Overall, the accuracy of the PHQ-9 diagnostic algorithms did not compare favorably to that of the PHQ-9 using the standard cutoff of ≥10 (sensitivity = 0.88, specificity = 0.86 compared to semi-structured diagnostic interview).

Whether or not screening should be implemented in practice is controversial. Screening in primary care settings is recommended by the USPSTF [8], but the Canadian Task Force on Preventive Health Care [9] and the United Kingdom National Screening Committee [10] both recommend against routine screening of people not reporting symptoms or suspected of possibly having depression. There are not any well-conducted randomized trials that have found that depression screening reduces depression symptoms or improves other patient-impor-

tant outcomes [6, 7, 9, 10, 52]. In this context, concerns have been raised about possible adverse effects for people screened, as well as the possibility of high false-positive rates, overdiagnosis, and substantial resource utilization and opportunity costs from screening [9, 10]. Well-conducted trials are needed to determine whether screening programs can be designed in a way that results in benefits to patients and minimizes harms and costs; concerns about false-positive screens and other negative implications of screening should be weighed against benefits demonstrated in clinical trials. Such trials can also be designed to determine what cutoff on the PHQ-9 may maximize benefits, if any, from screening and minimize harms. The standard cutoff for the PHQ-9 was set to maximize combined sensitivity and specificity, but that may not maximize clinical utility. Ideally, trials would be sufficiently large to compare benefits and harms from screening across different possible cutoff PHQ-9 thresholds. It is possible that further work on the measurement properties and scoring of the PHQ-9, such as with Rasch or Mokken analyses, may facilitate this also [53].

Beyond screening, the PHQ-9 diagnostic algorithm was designed to replicate DSM diagnostic criteria for major depression [11–13], and some authors have suggested that the PHQ-9 diagnostic algorithm could be used to diagnose depression and make treatment decisions for individual patients [11, 54, 55]. Although the PHQ-9 includes the same symptoms evaluated in assessing DSM major depression, it does not include all components of a diagnostic interview, including an assessment of functional impairment, investigation of nonpsychiatric medical conditions that can cause similar symptoms, or historical information necessary for differential diagnosis [3–5]. Thus, while the PHQ-9 may be used to solicit symptoms as part of a clinical assessment, it should not be used on a stand-alone basis for diagnosis; the present study showed that it would fail to diagnose approximately 40% of patients who meet diagnostic criteria for major depression.

We know of only one other self-report tool, the Major Depression Inventory (MDI) that, like the PHQ-9, was developed to be used as a summed score severity scale, as well as to include items that reflect standard diagnostic criteria [56–58]. Unlike the PHQ-9, though, the MDI was designed to capture both DSM and ICD criteria for major depression. Validation studies of the MDI, however, have been conducted in samples of people suspected of having depression or diagnosed with a depressive disorder [56–58], which limits comparability of results to those of the PHQ-9 from the present study. Thus, it is not clear wheth-

er the finding from the PHQ-9 that a cutoff threshold approach for screening provides a better balance of sensitivity and specificity would apply to the MDI.

This was the first study to use IPDMA to assess the accuracy of the PHQ-9 diagnostic algorithm approach for screening. Strengths include the large sample size, the ability to include results from studies with primary data rather than just those that published aggregate results, the ability to examine participant subgroups, and the ability to assess accuracy separately across reference standards, which had not been done previously. There are limitations to consider. First, we were unable to include primary data from 18 of 72 identified eligible data sets (25% of studies; 23% of participant data). Second, there was substantial heterogeneity across studies, although it did improve in some instances when subgroups were considered. There were not sufficient data to conduct subgroup analyses based on specific medical comorbidities or cultural aspects such as country or language. However, this was the first study of the PHQ-9 algorithm to compare participant subgroups based on age, sex, and country human development index. Third, we categorized studies based on the diagnostic interview administered, but interviews are sometimes adapted and may not always be used in the way that they were originally designed. Although we coded for interviewer qualification for all semi-structured interviews as part of our QUADAS-2 rating, two studies used interviewers who did not meet typical standards, and approximately half of studies were rated as unclear on this item. Finally, our study only addressed using the PHQ-9 for screening and not for other purposes, such as case finding or tracking treatment progress. We do not know of evidence on using the PHQ-9 for case finding among those already suspected of having depression, although others have examined this with other assessment tools [59, 60].

Conclusions

Diagnostic accuracy, or the ability to discriminate between people with and without a condition, is a core clinimetric criterion for evaluating the usefulness of a scale [3–5]. The results of the present study, in combination with those of a previous IPDMA, show that the PHQ-9 score threshold approach provides more desirable combinations of sensitivity and specificity across different cutoffs than the algorithm approach for screening and provides the flexibility to select a cutoff that would provide the preferred combination of sensitivity and specific-

ity. The algorithm approach may be attractive because it allows mapping of symptoms onto DSM diagnostic criteria and may be useful to provide information for an integrated mental health assessment. The PHQ-9 algorithms, however, are not sufficiently accurate to use exclusively for diagnosis, and empirical evidence also suggests that the algorithms do not perform as well as a score-based cutoff threshold approach for screening. Thus, the cutoff threshold approach is advised for use in clinical trials or if used in clinical practice. Even the cutoff approach, however, has limitations in that it crudely dichotomizes patients as positive or negative screens based on a single threshold with all symptom items counted equally. A risk-modeling approach could be used to generate individualized probabilities that a patient has major depression based on actual screening tool scores (rather than a dichotomous classification) and patient characteristics and could also weight responses for each PHQ-9 item differently. Ideally, to do this with acceptable precision, an even larger data set than used in the present study would be needed. Our team is working to compile such a data set, and we hope that this will be possible in the next few years.

Statement of Ethics

The authors have no ethical conflicts to disclose.

Disclosure Statement

Funding Sources

Author Contributions

C.H., B. Levis, J.B., P.C., S.G., J.P.A.I., L.A.K., D.M., S.B.P., I.S., R.J.S., R.C.Z., B.D.T., and A. Benedetti were responsible for the study conception and design. J.B. and L.A.K. designed and conducted database searches to identify eligible studies. D.H.A., B.A., L.A., H.R.B., M.B., A. Beraldi, C.H.B., H.B., P.B., G.C., M.H.C., J.C.N.C., R.C., K.C., Y.C., J.M.G., J.R.F., F.H.F., D.F., B.G., F.G.S., C.G.G., B.J.H., J.H., P.A.H., M. Härter, U.H., L.H., S.E.H., M. Hudson, T.H., M. Inagaki, K.I., N.J., M.E.K., K.M.K., Y.K., F.L., S.L., M.L., S.R.L., B. Löwe, L.M., A.M., S.M., T.N.M., K.M., F.L.O., V.P., B.W.P., P.P., A.P., K.R., A.G.R., I.S.S., J.S., A. Sidebottom, A. Simning, L.S., S.S., P.L.L.T., A.T., H.C.W., J.W., M.A.W., K.W., K.A.W., M.Y., and B.D.T. were responsible for collection of primary data included in this study. C.H., B. Levis, K.E.R., N.S., A.W.L., M.A., D.B.R., A.K., Y.W., Y.S., M. Imran, and B.D.T. contributed to data extraction and coding for the meta-analysis. C.H., B. Levis, B.D.T., and A.Benedetti contributed to the data analysis and interpretation. C.H., B. Levis, B.D.T., and A. Benedetti contributed to drafting the manuscript. All authors provided a critical review and approved the final paper. B.D.T. and A. Benedetti are the guarantors.

References

1 Raffle A, Gray M. Screening: evidence and practice. London, UK: Oxford University Press; 2007. https://doi.org/10.1093/acprof:oso/9780199214495.001.0001.

2 Wilson JM, Jungner G. Principles and practices of screening for disease. Geneva: World Health Organization; 1968.

3 Fava GA, Tomba E, Sonino N. Clinimetrics: the science of clinical measurements. Int J Clin Pract. 2012 Jan;66(1):11–5.

4 Fava GA, Carrozzino D, Lindberg L, Tomba E. The clinimetric approach to psychological assessment: a tribute to Per Bech, MD (1942-2018). Psychother Psychosom. 2018;87(6): 321–6.

5 Tomba E, Bech P. Clinimetrics and clinical psychometrics: macro- and micro-analysis. Psychother Psychosom. 2012;81(6):333–43.

6 Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care? BMJ. 2014 Feb 4; 348:g1253.

7 Thombs BD, Coyne JC, Cuijpers P, de Jonge P, Gilbody S, Ioannidis JP, et al. Rethinking recommendations for screening for depression in primary care. CMAJ. 2012 Mar;184(4): 413–8.

8 Siu AL, Bibbins-Domingo K, Grossman DC, Baumann LC, Davidson KW, Ebell M, et al.; US Preventive Services Task Force (USPSTF). Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. JAMA. 2016 Jan;315(4):380–7.

9 Joffres M, Jaramillo A, Dickinson J, Lewin G, Pottie K, Shaw E, et al.; Canadian Task Force on Preventive Health Care. Recommendations on screening for depression in adults. CMAJ. 2013 Jun;185(9):775–82.

10 Allaby M. Screening for depression: A report for the UK National Screening Committee (Revised report). London, United Kingdom: UK National Screening Committee; 2010.

11 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001 Sep;16(9): 606–13.

12 Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann. 2002;32(9):1–7.

13 Spitzer RL, Kroenke K, Williams JB; Primary Care Evaluation of Mental Disorders. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA. 1999 Nov;282(18):1737–44.

14 Maurer DM, Raymond TJ, Davis BN. Depression: screening and Diagnosis. Am Fam Physician. 2018 Oct;98(8):508–15.

15 American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-III. 3rd ed, revised. Washington: American Psychiatric Association; 1987.

16 American Psychiatric Association. Diagnostic and statistical manual of mental disorders:

DSM-IV. 4th ed. Washington: American Psychiatric Association; 1994.

17 American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV. 4th ed, text revised. Washington: American Psychiatric Association; 2000.

18 American Academy of Family Physicians. Depression [cited 2019 Feb 20]. Avaliable from: https://www.aafp.org/patient-care/clinical-recommendations/all/depression.html

19 Lichtman JH, Bigger JT Jr, Blumenthal JA, Frasure-Smith N, Kaufmann PG, Lespérance F, et al. Depression and coronary heart disease: recommendations for screening, referral, and treatment: a science advisory from the American Heart Association Prevention Committee of the Council on Cardiovascular Nursing, Council on Clinical Cardiology, Council on Epidemiology and Prevention, and Interdisciplinary Council on Quality of Care and Outcomes Research: endorsed by the American Psychiatric Association. Circulation. 2008 Oct;118(17):1768–75.

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

21 Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen Hosp Psychiatry. 2015 Jan-Feb;37(1):67–75.

22 Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. Gen Hosp Psychiatry. 2015 Nov-Dec;37(6):567–76.

23 Thombs BD, Arthurs E, El-Baalbaki G, Meijer A, Ziegelstein R, Steele R. Risk of bias from inclusion of already diagnosed or treated patients in diagnostic accuracy studies of depression screening tools: A systematic review. BMJ. 2011;343:d4825.

24 Rice DB, Thombs BD. Risk of bias from inclusion of currently diagnosed or treated patients in studies of depression screening tool accuracy: A cross-sectional analysis of recently published primary studies and meta-analyses. PLoS One. 2016 Feb;11(2):e0150067.

25 Levis B, Benedetti A, Riehm KE, Saadat N, Levis AW, Azar M, et al. Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. Br J Psychiatry. 2018 Jun; 212(6):377–85.

26 Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010 Feb; 340 feb05 1:c221.

27 Thombs BD, Benedetti A, Kloda LA, Levis B, Nicolau I, Cuijpers P, et al. 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. Syst Rev. 2014 Oct;3(1):124.

28 McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al.; the PRISMA-DTA Group. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA Statement. JAMA. 2018 Jan; 319(4):388–96.

29 Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA. 2015 Apr; 313(16):1657–65.

30 World Health Organization. The ICD-10 classifications of mental and behavioural disorder. Clinical descriptions and diagnostic guidelines. Geneva: World Health Organization; 1992.

31 PRESS – Peer Review of Electronic Search Strategies. 2015 guideline explanation and elaboration (PRESS E&E). Ottawa: CADTH; 2016 Jan.

32 United Nations. International human development indicators [cited 2019 Feb 20]. Available from: http://hdr.undp.org/en/countries

33 Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al.; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct;155(8):529–36.

34 First MB. Structured clinical interview for the DSM (SCID). John Wiley & Sons, Inc.; 1995.

35 World Health Organization. Schedules for Clinical Assessment in Neuropsychiatry: manual. Arlington: American Psychiatric Publishing; 1994.

36 Freedland KE, Skala JA, Carney RM, Raczynski JM, Taylor CB, Mendes de Leon CF, et al. The Depression Interview and Structured Hamilton (DISH): rationale, development, characteristics, and clinical validity. Psychosom Med. 2002 Nov-Dec;64(6):897–905.

37 Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Arch Gen Psychiatry. 1988 Dec;45(12):1069–77.

38 Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. Psychol Med. 1992 May;22(2):465–86.

39 Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. Arch Gen Psychiatry. 1981 Apr;38(4):381–9.

40 Lecrubier Y, Sheehan DV, Weiller E, Amorim P, Bonora I, Harnett Sheehan K, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. Eur Psychiatry. 1997;12(5): 224–31.

41 Sheehan DV, Lecrubier Y, Sheehan KH, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. Eur Psychiatry. 1997;12(5):232–41.

42 Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). Psychol Med. 2001 Aug;31(6):1001–13.

43 Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. Psychol Med. 1999 Sep;29(5):1013–20.

44 Nosen E, Woody SR. Chapter 8: diagnostic assessment in research. In: McKay D, editor. Handbook of research methods in abnormal and clinical psychology. Thousand Oaks: Sage; 2008.

45 Kurdyak PA, Gnam WH. Small signal, big noise: performance of the CIDI depression module. Can J Psychiatry. 2005 Nov;50(13): 851–6.

46 Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. Stat Med. 2008 Dec;27(29): 6111–36.

47 van der Leeden R, Busing FM, Meijer E. Bootstrap methods for two-level models. Technical Report PRM 97-04. Leiden, The Netherlands: Leiden University, Department of Psychology; 1997.

48 van der Leeden R, Meijer E, Busing FMTA. Chapter 11: resampling multilevel models. In: Leeuw J, Meijer E, editors. Handbook of multilevel analysis. New York: Springer; 2008. pp. 401–33.

49 Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002 Jun;21(11):1539–58.

50 Lambert SD, Clover K, Pallant JF, Britton B, King MT, Mitchell AJ, et al. Making sense of variations in prevalence estimates of depression in Cancer: a co-calibration of commonly used depression scales using Rasch analysis. J Natl Compr Canc Netw. 2015 Oct;13(10): 1203–11.

51 Amoozegar F, Patten SB, Becker WJ, Bulloch AG, Fiest KM, Davenport WJ, et al. The prevalence of depression and the accuracy of depression screening tools in migraine patients. Gen Hosp Psychiatry. 2017 Sep;48:25–31.

52 Thombs BD, Ziegelstein RC, Roseman M, Kloda LA, Ioannidis JP. There are no randomized controlled trials that support the United States Preventive Services Task Force Guideline on screening for depression in primary care: a systematic review. BMC Med. 2014 Jan;12(1):13.

53 Christensen KS, Oernboel E, Zatzick D, Russo J. Screening for depression: rasch analysis of the structural validity of the PHQ-9 in acutely injured trauma survivors. J Psychosom Res. 2017 Jun;97:18–22.

54 Dejesus RS, Vickers KS, Melin GJ, Williams MD. A system-based approach to depression management in primary care using the Patient Health Questionnaire-9. Mayo Clin Proc. 2007 Nov;82(11):1395–402.

55 Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care. 2003 Nov;41(11):1284–92.

56 Bech P, Rasmussen NA, Olsen LR, Noerholm V, Abildgaard W. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. J Affect Disord. 2001 Oct; 66(2-3):159–64.

57 Bech P, Timmerby N, Martiny K, Lunde M, Soendergaard S. Psychometric evaluation of the Major Depression Inventory (MDI) as depression severity scale using the LEAD (Longitudinal Expert Assessment of All Data) as index of validity. BMC Psychiatry. 2015 Aug;15(1):190.

58 Bech P, Christensen EM, Vinberg M, Østergaard SD, Martiny K, Kessing LV. The performance of the revised Major Depression Inventory for self-reported severity of depression—implications for the DSM-5 and ICD-11. Psychother Psychosom. 2013;82(3):187–8.

59 Christensen KS, Sokolowski I, Olesen F. Casefinding and risk-group screening for depression in primary care. Scand J Prim Health Care. 2011 Jun;29(2):80–4.

60 Nielsen MG, Ørnbøl E, Bech P, Vestergaard M, Christensen KS. The criterion validity of the web-based Major Depression Inventory when used on clinical suspicion of depression in primary care. Clin Epidemiol. 2017 Jul;9: 355–65.