This is the accepted manuscript version of an article published by S. Karger AG in [The accuracy of the Patient Health Questionaire-9 algorithm for screening to detect major depression: an individual participant data meta-analysis. Psychother Psychosom. 2020;89(1):25-37]

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

Chen He[1,2]; Brooke Levis[1,2]; Kira E. Riehm[1]; Nazanin Saadat[1]; Alexander W. Levis[1,2]; Marleine Azar[1,2]; Danielle B. Rice[1,3]; Ankur Krishnan[1]; Yin Wu[1,2,4] ; Ying Sun[1]; Mahrukh Imran[1]; Jill Boruff[5]; Pim Cuijpers[6], PhD; Simon Gilbody[7]; John P.A. Ioannidis[8]; Lorie A. Kloda[9]; Dean McMillan[7]; Scott B. Patten[10,11]; Ian Shrier[1,2]; Roy C. Ziegelstein[12]; Dickens H. Akena[13]; Bruce Arroll[14]; Liat Ayalon[15]; Hamid R. Baradaran[16,18]; Murray Baron[1,17]; Anna Beraldi[19]; Charles H. Bombardier[20]; Peter Butterworth[21,22]; Gregory Carter[23]; Marcos H. Chagas[24]; Juliana C. N. Chan[25,26,27]; Rushina Cholera[28]; Kerrie Clover[29,30]; Yeates Conwell[31]; Janneke M. de Man-van Ginkel[32]; Jesse R. Fann[33]; Felix H. Fischer[34]; Daniel Fung[35,36,37,38]; Bizu Gelaye[39]; Felicity Goodyear-Smith[14]; Catherine G. Greeno[40]; Brian J. Hall[41,42]; Patricia A. Harrison[43]; Martin Härter[44]; Ulrich Hegerl[45]; Leanne Hides[46]; Stevan E. Hobfoll[47]; Marie Hudson[1,17]; Thomas Hyphantis[48]; Masatoshi Inagaki[49]; Khalida Ismail[50]; Nathalie Jetté[10,11,51]; Mohammad E. Khamseh[16]; Kim M. Kiely[52,53]; Yunxin Kwan[54]; Femke Lamers[55]; Shen-Ing Liu[38,56,57,58]; Manote Lotrakul[59]; Sonia R. Loureiro[24]; Bernd Löwe[60]; Laura Marsh[61]; Anthony McGuire[62]; Sherina Mohd-Sidik[63]; Tiago N. Munhoz[64]; Kumiko Muramatsu[65]; Flávia L. Osório[24,66]; Vikram Patel[67,68]; Brian W. Pence[69]; Philippe Persoons[70,71]; Angelo Picardi[72]; Katrin Reuter[73]; Alasdair G. Rooney[74]; Iná S. Santos[64]; Juwita Shaaban[75]; Abbey Sidebottom[76]; Adam Simning[31]; Lesley Stafford[77,78]; Sharon Sung[35.38]; Pei Lin Lynnette Tan[54]; Alyna Turner[79,80]; Henk C. van Weert[81]; Jennifer White[82]; Mary A. Whooley[83,84,85]; Kirsty Winkley[86]; Mitsuhiko Yamada[87]; Brett D. Thombs[1,2,3,4,17,88]; Andrea Benedetti[2,17].

Author Affiliations:

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

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

[5] Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montreal, Quebec, Canada

[6] Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit, Amsterdam, the Netherlands

[7] Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK

[8] Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA

[9] Library, Concordia University, Montréal, Québec, Canada

[10] Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada

[11] Hotchkiss Brain Institute and O'Brien Institute for Public Health, University of Calgary, Calgary, Alberta, Canada

[12] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[13] Department of Psychiatry, Makerere University College of Health Sciences, Kampala, Uganda

[14] Department of General Practice and Primary Health Care, University of Auckland, New Zealand

[15] Louis and Gabi Weisfeld School of Social Work, Bar Ilan University, Ramat Gan, Israel

[16] Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran

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

[18] Ageing Clinical & Experimental Research Team, Institute of Applied Health Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, Scotland, UK

[19] Kbo-Lech-Mangfall-Klinik Garmisch-Partenkirchen, Klinik für Psychiatrie, Psychotherapie & Psychosomatik, Lehrkrankenhaus der Technischen Universität München, Munich, Germany

[20] Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA

[21] Centre for Research on Ageing, Health and Wellbeing, Research School of Population Health, The Australian National University, Canberra, Australia

[22] Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Melbourne, Australia

[23] Centre for Brain and Mental Health Research, University of Newcastle, New South Wales, Australia

[24] Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil

[25] Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China

[26] Asia Diabetes Foundation, Prince of Wales Hospital, Hong Kong Special Administrative Region, China

[27] Hong Kong Institute of Diabetes and Obesity, Hong Kong Special Administrative Region, China

[28] Department of Pediatrics, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA

[29] Centre for Brain and Mental Health Research, University of Newcastle, New South Wales, Australia

[30] Psycho-Oncology Service, Calvary Mater Newcastle, New South Wales, Australia

[31] Department of Psychiatry, University of Rochester Medical Center, Rochester, New York, USA

[32] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[33] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA

[34] Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Germany

[35] Department of Child & Adolescent Psychiatry, Institute of Mental Health, Singapore

[36] Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[37] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

[38] Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore

[39] Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

[40] School of Social Work, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[41] Global and Community Mental Health Research Group, Department of Psychology, Faculty of Social Sciences, University of Macau, Macau Special Administrative Region, China

[42] Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[43] City of Minneapolis Health Department, Minneapolis, Minnesota, USA

[44] Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[45] Department of Psychiatry, Psychosomatics and Psychotherapy, Goethe-Universität Frankfurt, German Depression Foundation, Frankfurt, Germany

[46] School of Psychology, University of Queensland, Brisbane, Queensland, Australia

[47] STAR-Stress, Anxiety, and Resilience Consultants, Chicago, Illinois, USA

[48] Faculty of Medicine, School of Health Sciences, University of Ioannina, Ioannina, Greece

[49] Department of Psychiatry, Faculty of Medicine, Shimane University, Shimane, Japan

[50] Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neurosciences, King's College London Weston Education Centre, London, UK

[51] Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[52] School of Psychology, University of New South Wales, Sydney, Australia

[53] Neuroscience Research Australia, Sydney, Australia

[54] Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore

[55] Department of Psychiatry, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit, Amsterdam, the Netherlands

[56] Department of Psychiatry, Mackay Memorial Hospital, Taipei, Taiwan

[57] Department of Medical Research, Mackay Memorial Hospital, Taipei, Taiwan

[58] Department of Medicine, Mackay Medical College, Taipei, Taiwan

[59] Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[60] Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[61] Baylor College of Medicine, Houston and Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas, USA

[62] Department of Nursing, St. Joseph's College, Standish, Maine, USA

[63] Cancer Resource & Education Centre, and Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

[64] Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil

[65] Department of Clinical Psychology, Graduate School of Niigata Seiryo University, Niigata, Japan

[66] National Institute of Science and Technology, Translational Medicine, Ribeirão Preto, Brazil

[67] Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, USA

[68] Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[69] Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[70] Department of Adult Psychiatry, University Hospitals Leuven, Leuven, Belgium

[71] Department of Neurosciences, Katholieke Universiteit Leuven, Leuven, Belgium

[72] Centre for Behavioural Sciences and Mental Health, Italian National Institute of Health, Rome, Italy

[73] Practice for Psychotherapy and Psycho-oncology, Freiburg, Germany

[74] Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburg, Edinburgh, Scotland, UK

[75] Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia

[76] Allina Health, Minneapolis, Minnesota, USA

[77] Centre for Women's Mental Health, Royal Women's Hospital, Parkville, Australia

[78] Melbourne School of Psychological Sciences, University of Melbourne, Australia

[79] School of Medicine and Public Health, University of Newcastle, New South Wales, Newcastle, Australia

[80] IMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, Victoria, Australia

[81] Department of General Practice, Amsterdam Institute for General Practice and Public Health, Amsterdam University Medical Centers, Amsterdam, the Netherlands

[82] Monash University, Melbourne, Australia

[83] Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA

[84] Department of Medicine, Veterans Affairs Medical Center, San Francisco, California, USA

[85] Department of Medicine, University of California San Francisco, San Francisco, California, USA

[86] Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, UK

[87] Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Ogawa-Higashi, Kodaira, Tokyo, Japan

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

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

Corresponding Authors:

Brett D. Thombs, PhD

Jewish General Hospital

4333 Cote Ste Catherine Road

Montreal, Quebec H3T 1E4

Tel: (514) 340-8222 ext. 25112

E-mail: brett.thombs@mcgill.ca

Andrea Benedetti, PhD

Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre

5252 Boulevard de Maisonneuve

Montréal, Quebec, H4A 3S5, Canada

Tel: (514) 934-1934 ext. 32161;

E-mail: andrea.benedetti@mcgill.ca

Keywords:

depression; diagnostic accuracy; meta-analysis; Patient Health Questionnaire-9; screening

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 individual participant data meta-analysis (IPDMA) 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 & Other Non-Indexed Citations, PsycINFO, Web of Science (January 1, 2000 – 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 to 0.24 lower compared to fully structured interviews and 0.06 to 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.

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 if 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 if 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 nine-item self-report questionnaire, is the most commonly used depression screening tool in primary care [14]. Its nine items align with the nine 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 two 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 five 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 one point for item nine (*thoughts that you would be better off dead or of hurting yourself in some way*) [13].

We 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 results 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

Datasets 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 two 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 two 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.

Datasets 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 if 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 & Other Non-Indexed Citations via Ovid, PsycINFO, and Web of Science (January 1, 2000 – February 7, 2015), using a peer-reviewed [31] search strategy (Supplementary Methods 1) 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 non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada).

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 datasets 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 (non-medical, 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 datasets 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 weight, 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 dataset with study-level data. We compared published participant characteristics and diagnostic accuracy results with results from raw datasets 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 over-inclusive [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 versus ≥ 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 1000 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% to 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 dataset used for the present evaluation of the diagnostic algorithms (N = 16,688).

All analyses were run in R (R version R 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 datasets (Supplementary Figure 1). In addition, authors of included studies contributed data from three unpublished studies (two subsequently published) [50, 51], for a total of 58 datasets of 72 identified eligible datasets (81%). Of these, four studies 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 Supplementary Table 1, Supplementary Table 2a, and 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 number (%) 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 to 0.64) and 0.95 (95% CI, 0.94 to 0.97) for the original algorithm and 0.61 (95% CI, 0.54 to 0.68) and 0.95 (0.93 to 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 Supplementary Figures 2a-2af and Supplementary Figures 3a-3af, with τ^2 and R values shown in 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 Supplementary Figures 4a-b and Supplementary Figures 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 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 subgroups based on age, sex, human development index, and care setting are shown in Supplementary Table 4, with forest plots shown in Supplementary Figures 2a-2af and Supplementary Figures 3a-3af, and τ^2 and R values shown in 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

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 dataset 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 versus 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-important 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 if 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 non-psychiatric 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 whether 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 datasets (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 specificity. 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 modelling 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 dataset than used in the present study would be needed. Our team is working to compile such a dataset, and we hope that this will be possible in the next years.

STATEMENTS

Acknowledgements: Not applicable

Statement of Ethics: The authors have no ethical conflicts to disclose

Disclosure Statement: All authors have completed the ICJME uniform disclosure form and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years with the following exceptions: Drs. Jetté and Patten declare that they received a grant, outside the submitted work, from the University of Calgary Hotchkiss Brain Institute, which was jointly funded by the Institute and Pfizer. Pfizer was the original sponsor of the development of the PHQ-9, which is now in the public domain. Dr. Chan is a steering committee member or consultant of Astra Zeneca, Bayer, Lilly, MSD and Pfizer. She has received sponsorships and honorarium for giving lectures and providing consultancy and her affiliated institution has received research grants from these companies. Dr. Hegerl declares that within the last three years, he was an advisory board member for Lundbeck and Servier; a consultant for Bayer Pharma; a speaker for Roche Pharma and Servier; and received personal fees from Janssen, all outside the submitted work. Dr. Inagaki declares that he has received a grant from Novartis Pharma, and personal fees from Meiji, Mochida, Takeda, Novartis, Yoshitomi, Pfizer, Eisai, Otsuka, MSD, Technomics, and Sumitomo Dainippon, all outside of the submitted work. All authors declare no other relationships or activities that could appear to have influenced the submitted work. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Funding Sources: This study was funded by the Canadian Institutes of Health Research (CIHR; KRS-134297, PCG-155468). Ms. Levis was supported by a CIHR Frederick Banting and Charles

Best Canada Graduate Scholarship doctoral award. Ms. Riehm and Ms. Saadat were supported by CIHR Frederick Banting and Charles Best Canadian Graduate Scholarships – Master's Awards. Mr. Levis and Ms. Azar were supported by FRQS Masters Training Awards. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. Dr. Wu was supported by an Utting Postdoctoral Fellowship from the Jewish General Hospital, Montreal, Quebec, Canada. Collection of data for the study by Arroll et al. was supported by a project grant from the Health Research Council of New Zealand. Data collection for the study by Ayalon et al. was supported from a grant from Lundbeck International. The primary study by Khamseh et al. was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary study by Bombardier et al. was supported by the Department of Education, National Institute on Disability and Rehabilitation Research, Spinal Cord Injury Model Systems: University of Washington (grant no. H133N060033), Baylor College of Medicine (grant no. H133N060003), and University of Michigan (grant no. H133N060032). Dr. Butterworth was supported by Australian Research Council Future Fellowship FT130101444. Dr. Cholera was supported by a United States National Institute of Mental Health (NIMH) grant (5F30MH096664), and the United States National Institutes of Health (NIH) Office of the Director, Fogarty International Center, Office of AIDS Research, National Cancer Center, National Heart, Blood, and Lung Institute, and the NIH Office of Research for Women's Health through the Fogarty Global Health Fellows Program Consortium (1R25TW00934001) and the American Recovery and Reinvestment Act. Dr. Conwell received support from NIMH (R24MH071604) and the Centers for Disease Control and Prevention (R49 CE002093). The primary studies by Amoozegar and by Fiest et al. were funded by the Alberta Health Services, the University of Calgary Faculty of Medicine, and the Hotchkiss Brain Institute. The primary study by Fischer et al. was funded by the German Federal Ministry of Education and Research (01GY1150). Data for the

primary study by Gelaye et al. was supported by grant from the NIH (T37 MD001449). Collection of data for the primary study by Gjerdingen et al. was supported by grants from the NIMH (R34 MH072925, K02 MH65919, P30 DK50456). The primary study by Eack et al. was funded by the NIMH (R24 MH56858). Collection of data provided by Drs. Härter and Reuter was supported by the Federal Ministry of Education and Research (grants No. 01 GD 9802/4 and 01 GD 0101) and by the Federation of German Pension Insurance Institute. Collection of data for the primary study by Hobfoll et al. was made possible in part from grants from NIMH (RO1 MH073687) and the Ohio Board of Regents. Dr. Hall received support from a grant awarded by the Research and Development Administration Office, University of Macau (MYRG2015-00109-FSS). The primary study by Hides et al. was funded by the Perpetual Trustees, Flora and Frank Leith Charitable Trust, Jack Brockhoff Foundation, Grosvenor Settlement, Sunshine Foundation and Danks Trust. The primary study by Henkel et al. was funded by the German Ministry of Research and Education. Data for the study by Razykov et al. was collected by the Canadian Scleroderma Research Group, which was funded by the CIHR (FRN 83518), the Scleroderma Society of Canada, the Scleroderma Society of Ontario, the Scleroderma Society of Saskatchewan, Sclérodermie Québec, the Cure Scleroderma Foundation, Inova Diagnostics Inc., Euroimmun, FRQS, the Canadian Arthritis Network, and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montreal, QC. Dr. Hudson was supported by a FRQS Senior Investigator Award. Collection of data for the primary study by Hyphantis et al. was supported by grant from the National Strategic Reference Framework, European Union, and the Greek Ministry of Education, Lifelong Learning and Religious Affairs (ARISTEIA-ABREVIATE, 1259). The primary study by Inagaki et al. was supported by the Ministry of Health, Labour and Welfare, Japan. Dr. Jetté was supported by a Canada Research Chair in Neurological Health Services Research. Collection of data for the

primary study by Kiely et al. was supported by National Health and Medical Research Council (grant number 1002160) and Safe Work Australia. Dr. Kiely was supported by funding from a Australian National Health and Medical Research Council fellowship (grant number 1088313). The primary study by Lamers et al. was funded by the Netherlands Organisation for Health Research and development (grant number 945-03-047). The primary study by Liu et al. was funded by a grant from the National Health Research Institute, Republic of China (NHRI-EX97-9706PI). The primary study by Lotrakul et al. was supported by the Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (grant number 49086). Dr. Bernd Löwe received research grants from Pfizer, Germany, and from the medical faculty of the University of Heidelberg, Germany (project 121/2000) for the study by Gräfe et al. The primary study by Mohd-Sidik et al. was funded under the Research University Grant Scheme from Universiti Putra Malaysia, Malaysia and the Postgraduate Research Student Support Accounts of the University of Auckland, New Zealand. The primary study by Santos et al. was funded by the National Program for Centers of Excellence (PRONEX/FAPERGS/CNPq, Brazil). The primary study by Muramatsu et al. was supported by an educational grant from Pfizer US Pharmaceutical Inc. Collection of primary data for the study by Dr. Pence was provided by NIMH (R34MH084673). The primary studies by Osório et al. were funded by Reitoria de Pesquisa da Universidade de São Paulo (grant number 09.1.01689.17.7) and Banco Santander (grant number 10.1.01232.17.9). Dr. Osório was supported by Productivity Grants (PQ-CNPq-2 -number 301321/2016-7). The primary study by Picardi et al. was supported by funds for current research from the Italian Ministry of Health. Dr. Persoons was supported by a grant from the Belgian Ministry of Public Health and Social Affairs and a restricted grant from Pfizer Belgium. Dr. Shaaban was supported by funding from Universiti Sains Malaysia. The primary study by Rooney et al. was funded by the United Kingdom National Health Service Lothian Neuro-Oncology

Endowment Fund. The primary study by Sidebottom et al. was funded by a grant from the United States Department of Health and Human Services, Health Resources and Services Administration (grant number R40MC07840). Simning et al.'s research was supported in part by grants from the NIH (T32 GM07356), Agency for Healthcare Research and Quality (R36 HS018246), NIMH (R24 MH071604), and the National Center for Research Resources (TL1 RR024135). Dr. Stafford received PhD scholarship funding from the University of Melbourne. Collection of data for the studies by Turner et al were funded by a bequest from Jennie Thomas through the Hunter Medical Research Institute. Collection of data for the primary study by Williams et al. was supported by a NIMH grant to Dr. Marsh (RO1-MH069666). The primary study by Thombs et al. was done with data from the Heart and Soul Study (PI Mary Whooley). The Heart and Soul Study was funded by the Department of Veterans Epidemiology Merit Review Program, the Department of Veterans Affairs Health Services Research and Development service, the National Heart Lung and Blood Institute (R01 HL079235), the American Federation for Aging Research, the Robert Wood Johnson Foundation, and the Ischemia Research and Education Foundation. Dr. Thombs was supported by an Investigator Award from the Arthritis Society. The primary study by Twist et al. was funded by the UK National Institute for Health Research under its Programme Grants for Applied Research Programme (grant reference number RP-PG-0606-1142). The study by Wittkampf et al. was funded by The Netherlands Organization for Health Research and Development (ZonMw) Mental Health Program (nos. 100.003.005 and 100.002.021) and the Academic Medical Center/University of Amsterdam. Collection of data for the primary study by Zhang et al. was supported by the European Foundation for Study of Diabetes, the Chinese Diabetes Society, Lilly Foundation, Asia Diabetes Foundation and Liao Wun Yuk Diabetes Memorial Fund. Drs. Thombs and Benedetti were

supported by Fonds de recherche du Québec - Santé (FRQS) researcher salary awards. No other authors reported funding for primary studies or for their work on the present study.

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

.

REFERENCES

1. Raffle A, Gray M. Screening: evidence and practice. London (UK): Oxford University Press; 2007.
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;66(1):11-15.
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):321-326.
5. Tomba E, Bech P. Clinimetrics and clinical psychometrics: macro- and micro-analysis. Psychother Psychosom. 2012;81(6):333-343.
6. Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care? BMJ. 2014;348:g1253.
7. Thombs BD, Coyne JC, Cuijpers P, et al. Rethinking recommendations for screening for depression in primary care. CMAJ. 2012;184(4):413-418.
8. Siu AL, and the US Preventive Services Task Force (USPSTF). Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. JAMA. 2016;315(4):380-387.
9. Joffres M, Jaramillo A, Dickinson J, et al. Recommendations on screening for depression in adults. CMAJ. 2013;185(9):775-782.
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;16(9):606-613.
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. 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;282(18):1737-1744.
14. Maurer DM, Raymond TJ, Davis BN. Depression: Screening and Diagnosis. Am Fam Physician. 2018;98(8):508-515.
15. Diagnostic and statistical manual of mental disorders: DSM-III 3rd ed, revised. Washington, DC: American Psychiatric Association 1987.
16. Diagnostic and statistical manual of mental disorders: DSM-IV 4th ed.Washington, DC: American Psychiatric Association 1994.
17. Diagnostic and statistical manual of mental disorders: DSM-IV 4th ed, text revised. Washington, DC: American Psychiatric Association 2000.
18. American Academy of Family Physicians [internet]. 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, 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;118(17):1768-1775.

20. Levis B, Benedetti A, Riehm KE, et al. The diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: An individual participant data meta-analysis. BMJ. 2019;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. General hospital psychiatry. 2015;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;37(6):567-576.
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;11(2):e0150067.
25. Levis B, Benedetti A, Riehm KE, et al. Probability of major depression diagnostic classification using semi-structured vs. fully structured diagnostic interviews. Br J Psychiatry. 2018;212(6):377–385.
26. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221.
27. Thombs BD, Benedetti A, Kloda LA, 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;3(1):124.

28. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. 2018;319(4):388-396.
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;313:1657-1665.
30. 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 [Internet]. International Human Development Indicators[cited 2019 Feb 20]. Avaliable from: http://hdr.undp.org/en/countries.
33. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536.
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. Amer Psychiatric Pub Inc. 1994.
36. Freedland KE, Skala JA, Carney RM, et al. The Depression Interview and Structured Hamilton (DISH): rationale, development, characteristics, and clinical validity. Psychosom Med. 2002;64(6):897-905.

37. Robins LN, Wing J, Wittchen HU, 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. 1988:45(12):1069-1077.
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;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;38(4):381-389.
40. Lecrubier Y, Sheehan DV, Weiller E, 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):224-231.
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):232-241.
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;31(6):1001-1013.
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;29(5):1013-1020.
44. Nosen E, Woody SR. Chapter 8: Diagnostic Assessment in Research. In, McKay D. Handbook of research methods in abnormal and clinical psychology. Sage; 2008.

45. Kurdyak PA, Gnam WH. Small signal, big noise: performance of the CIDI depression module. Can J Psychiatry. 2005;50(13):851-856.
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;27(29):6111-6136.
47. van der Leeden R, Busing FMTA, Meijer E. Bootstrap methods for two-level models. Technical Report PRM 97-04, Leiden University, Department of Psychology, Leiden, The Netherlands, 1997.
48. van der Leeden R, Meijer E, Busing FMTA. Chapter 11: Resampling multilevel models. In: Leeuw J, Meijer E, eds. Handbook of multilevel analysis New York, NY: Springer; 2008:401-433.
49. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-1558.
50. Lambert SD, Clover K, Pallant JF, 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;13(10):1203-11.
51. Amoozegar F, Patten SB, Becker WJ, et al. The prevalence of depression and the accuracy of depression screening tools in migraine patients. Gen Hosp Psychiatry. 2017;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;12:13.
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;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;82(11):1395-1402.

55. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2. Validity of a two-item depression screener. Medical Care. 2003;41(11):1284-1292.

56. Bech P, Rasmussen N-A, Raabæk Olsen L, 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;66(2-3):159-164.

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;15:190.

58. Bech P, Christensen EM, Vinberg M, et al. 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):187-188.

59. Christensen KS, Sokolowski I, Olesen F. Case-finding and risk-group screening for depression in primary care. Scand J Prim Health Care. 2011;29(2):80–84.

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;9:355–365.