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

Authors:

Brooke Levis, Andrea Benedetti, Brett D. Thombs and the DEPRESsion Screening Data (DEPRESSD) Collaboration.

DEPRESSD Collaboration:

Kira E. Riehm, Nazanin Saadat, Alexander W. Levis, Marleine Azar, Danielle B. Rice, Matthew J. Chiovitti, Tatiana A. Sanchez, Jill Boruff, Pim Cuijpers, Simon Gilbody, John P.A. Ioannidis, Lorie A. Kloda, Dean McMillan, Scott B. Patten, Ian Shrier, Roy C. Ziegelstein, Dickens H. Akena, Bruce Arroll, Liat Ayalon, Hamid R. Baradaran, Murray Baron, Charles H. Bombardier, Peter Butterworth, Gregory Carter, Marcos H. Chagas, Juliana C. N. Chan, Kerrie Clover, Yeates Conwell, Janneke M. de Man-van Ginkel, Jaime Delgadillo, Jesse R. Fann, Felix H. Fischer, Daniel Fung, Bizu Gelaye, Felicity Goodyear-Smith, Catherine G. Greeno, Brian J. Hall, John Hambridge, Patricia A. Harrison, Martin Härter, Ulrich Hegerl, Leanne Hides, Stevan E. Hobfoll, Marie Hudson, Masatoshi Inagaki, Khalida Ismail, Nathalie Jetté, Mohammad E. Khamseh, Kim M. Kiely, Yunxin Kwan, Shen-Ing Liu, Manote Lotrakul, Sonia R. Loureiro, Bernd Löwe, Laura Marsh, Anthony McGuire, Sherina Mohd Sidik, Tiago N. Munhoz, Kumiko Muramatsu, Flávia L. Osório, Vikram Patel, Brian W. Pence, Philippe Persoons, Angelo Picardi, Katrin Reuter, Alasdair G. Rooney, Iná S. Santos, Juwita Shaaban, Abbey Sidebottom, Adam Simning, Lesley Stafford, Sharon C. Sung, Pei Lin Lynnette Tan, Alyna Turner, Christina M. van der Feltz-Cornelis, Henk C. van

Weert, Paul A. Vöhringer, Jennifer White, Mary A. Whooley, Kirsty Winkley, Mitsuhiko Yamada, Yuying Zhang,

Affiliations:

Lady Davis Institute for Medical Research, Jewish General Hospital and McGill University, Montréal, Québec, Canada

Brooke Levis (doctoral student)

Kira E. Riehm (research assistant)

Nazanin Saadat (research assistant)

Alexander W. Levis (masters student)

Marleine Azar (masters student)

Danielle B. Rice (doctoral student)

Matthew J. Chiovitti (research assistant)

Tatiana A. Sanchez (research assistant)

Ian Shrier (sport medicine physician)

Murray Baron (rheumatologist)

Marie Hudson (rheumatologist)

Brett D. Thombs (professor)

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

Andrea Benedetti (associate professor)

Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montréal, Québec, Canada
Jill Boruff (associate librarian)

Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit, Amsterdam, the Netherlands
Pim Cuijpers (professor)

Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK
Simon Gilbody (professor)
Dean McMillan (reader)
Christina M. van der Feltz-Cornelis (professor)

Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA
John P.A. Ioannidis (professor)

Library, Concordia University, Montréal, Québec, Canada
Lorie A. Kloda (senior librarian)

Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada
Scott Patten (professor)

Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
Roy C. Ziegelstein (professor)

Department of Psychiatry, Makerere University College of Health Sciences, Kampala, Uganda
Dickens H. Akena (psychiatrist)

Department of General Practice and Primary Health Care, University of Auckland, New Zealand
Bruce Arroll (professor)
Felicity Goodyear-Smith (professor)

Louis and Gabi Weisfeld School of Social Work, Bar Ilan University, Ramat Gan, Israel
Liat Ayalon (professor)

Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran
Hamid R. Baradaran (professor)
Mohammad E. Khamseh (professor)

Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA
Charles H. Bombardier (professor)

Centre for Research on Ageing, Health and Wellbeing, Research School of Population Health, The Australian National University, Canberra, Australia
Kim M. Kiely (NHMRC Fellow)

Centre for Mental Health, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia
Peter Butterworth (professor)

Centre for Brain and Mental Health Research, University of Newcastle, New South Wales, Australia
Gregory Carter (conjoint professor)
Kerrie Clover (clinical psychologist)

Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil
Marcos H. Chagas (assistant professor)
Sonia R. Loureiro (professor)
Flávia L. Osório (teacher)

Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China
Juliana C. N. Chan (professor)
Yuying Zhang (researcher)

Psycho-Oncology Service, Calvary Mater Newcastle, New South Wales, Australia
Kerrie Clover (clinical psychologist)
Adam Simning (assistant professor)

Department of Psychiatry, University of Rochester Medical Center, Rochester, New York, USA
Yeates Conwell (professor)

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University Utrecht, Utrecht, the Netherlands
Janneke M. de Man-van Ginkel (assistant professor)

Clinical Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, UK
Jaime Delgadillo (lecturer in clinical psychology)

Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA
Jesse R. Fann (professor)

Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany, Germany
Felix H. Fischer (research fellow)

Department of Child & Adolescent Psychiatry, Institute of Mental Health, Singapore
Daniel Fung (associate professor)

Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA
Bizu Gelaye (assistant professor)

School of Social Work, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
Catherine G. Greeno (associate professor)

Global and Community Mental Health Research Group, Department of Psychology, Faculty of Social Sciences, University of Macau, Macau Special Administrative Region, China
Brian J. Hall (associate professor)

Liaison Psychiatry Department, John Hunter Hospital, Newcastle, Australia
John Hambridge (clinical psychologist)

City of Minneapolis Health Department, Minneapolis, Minnesota, USA
Patricia A. Harrison (director of research and evaluation)

Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
Martin Härter (professor)

Department of Psychiatry and Psychotherapy, German Depression Foundation, Leipzig, Germany

Ulrich Hegerl (professor)

School of Psychology, University of Queensland, Brisbane, Queensland, Australia

Leanne Hides (professor)

STAR-Stress, Anxiety & Resilience Consultants, Chicago, Illinois, USA

Stevan E. Hobfoll (managing member)

Department of Psychiatry, Faculty of Medicine, Shimane University, Shimane, Japan

Masatoshi Inagaki (professor)

Department of Psychological Medicine, Institute of Psychiatry, Psychology and

Neurosciences, King's College London Weston Education Centre, London, UK

Khalida Ismail (professor)

Department of Neurology, Ichan School of Medicine at Mount Sinai, New York, New York, USA

Nathalie Jetté (professor)

Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore

Yunxin Kwan(psychiatrist)

Pei Lin Lynnette Tan, MMed (psychiatrist)

Department of Psychiatry, Mackay Memorial Hospital, Taipei, Taiwan
Shen-Ing Liu (professor)

Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand
Manote Lotrakul (professor)

Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
Bernd Löwe (professor)

Baylor College of Medicine, Houston and Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas, USA
Laura Marsh (professor)

Department of Nursing, St. Joseph's College, Standish, Maine, USA
Anthony McGuire (professor)

Cancer Resource & Education Centre, and Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
Sherina Mohd Sidik (professor)

Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil
Tiago N. Munhoz (professor)
Iná S. Santos (professor)

Department of Clinical Psychology, Graduate School of Niigata Seiryo University, Niigata, Japan
Kumiko Muramatsu (psychiatrist)

Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, USA
Vikram Patel (professor)

Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
Brian W. Pence (associate professor)

Mind-Body Research, Department of Neurosciences, Katholieke Universiteit Leuven, Leuven, Belgium
Philippe Persoons (assistant professor)

Centre for Behavioural Sciences and Mental Health, Italian National Institute of Health, Rome, Italy
Angelo Picardi (senior researcher)

Group Practice for Psychotherapy and Psycho-oncology, Freiburg, Germany

Katrin Reuter (psychologist)

Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburg, Edinburgh, Scotland, UK
Alasdair G. Rooney (physician)

Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia
Juwita Shaaban, MMed (family medicine specialist)

Allina Health, Minneapolis, Minnesota, USA
Abbey Sidebottom (epidemiologist)

Melbourne School of Psychological Sciences, University of Melbourne, Australia
Lesley Stafford (associate professor)

Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore
Sharon C. Sung (assistant professor)

IMPACT Strategic Research Centre, School of Medicine, Deakin University, Geelong, Victoria, Australia
Alyna Turner (senior lecturer)

Department of General Practice, Amsterdam Institute for General Practice and Public Health, Amsterdam University Medical Centers, Location AMC, Amsterdam, the Netherlands
Henk C. van Weert (professor)

Millennium Institute for Depression and Personality Research (MIDAP), Ministry of Economy, Macul, Santiago, Chile
Paul A. Vöhringer (adjunct researcher)

Monash University, Melbourne, Australia
Jennifer White (research fellow)

Department of Medicine, Veterans Affairs Medical Center, San Francisco, California, USA
Mary A. Whooley (professor)

Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, Waterloo Road, London, UK
Kirsty Winkley (reader)

Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Ogawa-Higashi, Kodaira, Tokyo, Japan
Mitsuhiko Yamada (director)

Corresponding author:

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; ORCID: 0000-0002-5644-8432

Word count: 4,835

ABSTRACT

Objective: The Patient Health Questionnaire-9 (PHQ-9) has been recommended for screening to identify patients with depression. Conventional meta-analyses have been limited by selective cutoff reporting in primary studies and have not examined accuracy for different reference standards or participant subgroups. This study aimed to determine PHQ-9 accuracy for detecting major depression using individual participant data meta-analysis (IPDMA).

Design: IPDMA.

Data sources: Medline, Medline In-Process & Other Non-Indexed Citations, PsycINFO, and Web of Science (January 2000-February 2015).

Eligibility criteria for selecting studies: Eligible studies compared PHQ-9 scores to major depression diagnoses from validated diagnostic interviews. Primary study data and study-level data extracted from primary reports were synthesized. For PHQ-9 cutoffs 5 to 15, bivariate random-effects meta-analysis was used to estimate pooled sensitivity and specificity, separately, among studies that used semi-structured diagnostic interviews, which are designed for clinician administration; fully structured interviews, which are designed for lay administration; and the Mini International Neuropsychiatric (MINI) diagnostic interviews, a brief fully structured interview. Sensitivity and specificity were examined among participant subgroups and, separately, using meta-regression, considering all subgroup variables in a single model.

Results: Data were obtained for 58 of 72 eligible studies (N participants = 17,357, N cases = 2,312). Combined sensitivity and specificity was maximized at a cutoff of ≥10 among studies using a semi-structured interview (N = 29 studies, 6,725 participants; sensitivity [95% CI] = 0.88 [0.83 to 0.92], specificity [95% CI] = 0.85 [0.82 to 0.88]). Across cutoffs 5 to 15, sensitivity with semi-structured interviews was 5 to 22% higher than for fully structured interviews (MINI excluded; N = 14 studies,

7,680 participants) and 2 to15% higher than for the MINI (N = 15 studies, 2,952 participants). Specificity was similar across diagnostic interviews. The PHQ-9 appears to be similarly sensitive but may be less specific for younger patients than for older patients; a cutoff of ≥10 can be used regardless of age. In studies conducted in primary care using a semi-structured interview (N = 9 studies, 3,163 participants), sensitivity [95% CI] = 0.94 [0.88 to 0.97] and specificity [95% CI] = 0.88 [0.79 to 0.93]).

Conclusions: PHQ-9 sensitivity compared to semi-structured diagnostic interviews was greater than in previous conventional meta-analyses that combined reference standards. A cutoff of ≥10 maximized combined sensitivity and specificity overall and for subgroups.

Funding: Canadian Institutes of Health Research (KRS-134297; PCG-155468).

Registration: PROSPERO (CRD42014010673).

Depression screening refers to the use of a depression screening questionnaire to identify patients who may have depression, but have not been identified. When screening programs are recommended, clinicians are advised to administer a depression symptom questionnaire and to use a pre-identified cutoff threshold to classify patients as having positive or negative screening results. Those with positive screening results can then be evaluated to determine if they have depression and, if appropriate, be offered treatment.[1,2]

The Patient Health Questionnaire-9 (PHQ-9)[3-5] is a nine-item questionnaire designed to screen for depression in primary care and other medical settings.[6,7] The standard cutoff for screening to identify possible major depression is ≥10,[3-7] which was established in the first study on the PHQ-9 (N total = 580, N major depression = 41).[3,5]

A conventional PHQ-9 meta-analysis from 2015 (N studies = 36, N participants = 21,292),[8] evaluated sensitivity and specificity for cutoffs 7 to15 by combining accuracy results for each cutoff that were published in included primary studies. Pooled sensitivity for the standard cutoff of 10 was 0.78 (95% confidence interval [CI] = 0.70 to 0.84), and pooled specificity was 0.87 (95% CI = 0.84 to 0.90). Incomplete reporting of results from cutoffs other than 10 in the primary studies that were included, however, resulted in cutoff ranges where sensitivity implausibly increased as cutoff scores increased.[8] This suggested possible selective cutoff reporting in some primary studies to maximize accuracy.[8,9] Additional limitations included the inability to assess differences across patient subgroups, since subgroup results were not reported in primary studies; the inability to exclude participants already diagnosed or being treated for depression, who would not be screened in practice, but were included in many primary studies;[10,11] and the combining of accuracy estimates without differentiating between reference standards.[12] Semi-structured diagnostic interviews (e.g., Structured Clinical Interview for DSM Disorders [SCID][13]) are intended to be conducted by

experienced diagnosticians and require clinical judgment. Fully structured interviews (e.g., Composite International Diagnostic Interview [CIDI][14]) are fully scripted and designed to be administered by lay interviewers in order to reduce the cost of employing trained clinical interviewers; they are intended to achieve a high level of standardization, but may sacrifice accuracy.[15-18] The Mini International Neuropsychiatric Interview (MINI) is fully structured, but was designed for very rapid administration and described by its authors as being over-inclusive as a result.[19,20] In a recent analysis, controlling for depressive symptom scores, we found that the MINI classified approximately twice as many participants with major depression as other fully structured interviews. Compared to semi-structured interviews, fully structured interviews (MINI excluded) classified more patients with low symptom scores but fewer patients with high symptom scores as having major depression.[12]

Individual participant data meta-analysis (IPDMA) involves a standard systematic review, then synthesis of participant-level data from primary studies rather than summary results from study reports.[21] Advantages include the ability to conduct subgroup analyses not reported in primary studies, the ability to report results from all relevant cutoffs from all included studies, and the ability to exclude already diagnosed or treated participants who would not be screened in practice.

The objectives of this study were to use IPDMA to evaluate the diagnostic accuracy of the PHQ-9 screening tool (1) among studies using semi-structured, fully structured (MINI excluded), and MINI diagnostic interviews as reference standards, separately, with priority given to semi-structured interview results; (2) among participants not currently diagnosed or receiving treatment for a mental health problem; and (3) among participant subgroups based on age, sex, country human development index, and recruitment setting.

METHOD

This IPDMA was registered in PROSPERO (CRD42014010673), a protocol was published,[22] and results were reported following PRISMA-DTA[23] and PRISMA-IPD[24] reporting guidelines.

Search strategy

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) on February 7, 2015, using a peer-reviewed[25] search strategy (eMethods1). The search was limited to the year 2000 forward because the PHQ-9 was published in 2001.[3] 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) for storing and tracking search results.

Identification of eligible studies

Datasets from articles in any language were eligible for inclusion 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 who were not recruited from youth or psychiatric settings or because they were identified as having symptoms of depression. We required the diagnostic interviews and PHQ-9 to be administered within two weeks of each other because Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Classification of Diseases (ICD) major depression diagnostic criteria specify that symptoms must have been present in the last two weeks. We excluded patients from psychiatric settings or those already identified as having symptoms of depression because screening is done to identify previously unrecognized cases.

Datasets where not all participants were eligible were included if primary data allowed selection of eligible participants. For defining major depression, we considered MDD or MDE based on the DSM or ICD. If more than one was reported, we prioritized MDE over MDD, since screening would attempt to detect depressive episodes and further interview would determine if the episode is related to MDD or bipolar disorder, and DSM over ICD. Across all studies, there were 23 discordant diagnoses depending on classification prioritization (0.1% of participants).

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 for which team members were fluent.

Data extraction, contribution and synthesis

Authors of eligible datasets were invited to contribute de-identified primary data. 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”, “high”, or “low-medium” development based on the United Nation’s human development index.[26] Participant-level data included age, sex, major depression status, current mental health diagnosis or treatment, and PHQ-9 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 provided weights. 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.

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 the original investigators.

Two investigators assessed risk of bias of included studies independently, based on the primary publications, using the Quality Assessment of Diagnostic Accuracy Studies-2 tool (QUADAS-2; eMethods2).[27] Discrepancies were resolved by consensus.

Statistical analyses

We conducted three main sets of analyses. First, we estimated sensitivity and specificity across PHQ-9 cutoffs 5 to 15 for studies with semi-structured (SCID[13], Schedules for Clinical Assessment in Neuropsychiatry[28], Depression Interview and Structured Hamilton[29]), fully structured (MINI excluded; CIDI[14], Clinical Interview Schedule-Revised[30], Diagnostic Interview Schedule[31]), and MINI[19,20] reference standards, separately. Second, for each reference standard category, we estimated sensitivity and specificity across PHQ-9 cutoffs for all participants from primary studies, as has been done in existing conventional meta-analyses and, separately, among only participants who could be confirmed as not currently diagnosed or receiving treatment for a mental health problem at the time of assessment. This was done because existing conventional meta-analyses have all been based on primary studies that generally do not exclude patients already diagnosed or receiving treatment. Since screening is done to identify previously unrecognized cases, however, those patients would not be screened in practice, and their inclusion in diagnostic accuracy studies could bias results.[10,11] Third, for each reference standard category, we estimated and compared sensitivity and specificity across PHQ-9 cutoffs among subgroups

based on age (<60 versus ≥60 years), sex, country human development index (very high; high; low-medium), and recruitment setting (non-medical; primary; inpatient specialty; outpatient specialty). Among studies that used the MINI, we combined inpatient and outpatient specialty care settings, as only one study included inpatient participants. In each subgroup analysis, we excluded primary studies with no major depression cases, as this did not allow application of the bivariate random effects model. This resulted in a maximum of 15 participants excluded from any subgroup analysis.

For each meta-analysis, for cutoffs 5 to 15 separately, bivariate random-effects models were fitted via Gauss-Hermite adaptive quadrature.[32] This 2-stage meta-analytic approach models 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.

To compare results across reference standards and other subgroups, we constructed empirical receiver operating characteristic (ROC) curves for each group based on the pooled sensitivity and specificity estimates and calculated areas under the curve (AUC). We estimated differences in sensitivity and specificity between subgroups at each cutoff by constructing confidence intervals for differences via the cluster bootstrap approach,[33,34] resampling at study and subject levels. For each comparison, we ran 1000 iterations of the bootstrap. We removed iterations that did not produce difference estimates for cutoffs 5 to 15 prior to determining confidence intervals and noted the number of iterations removed.

In addition to categorical subgroup analyses, we compared sensitivity and specificity across the different reference standards by conducting one-stage meta-regressions with interactions between reference standard category (reference category = semi-structured interviews) and accuracy

coefficients (logit(sensitivity) and logit(specificity)), and we compared results to those seen in the original two-stage bivariate random effects meta-analytic models. Additionally, within each reference standard category, we conducted one-stage meta-regressions where we interacted all subgrouping variables (age [measured continuously], sex [reference category = women], country human development index [reference category = very high] and participant recruitment setting [reference category = primary care]) with logit(sensitivity) and logit(specificity). Similar to our main subgroup analyses, we once again determined which significant interactions replicated across all three reference standard categories. For subgrouping variables that were significantly associated with sensitivity or specificity coefficients for all three reference standard categories for all or most cutoffs in the main one-stage meta-regression, we conducted additional one-stage meta-regressions to produce accuracy estimates for the subgroups of interest, and we compared these results to those seen in the original two-stage bivariate random effects meta-analytic models. Although age was included as a continuous variable in the main meta-regression, we again dichotomized it (<60 versus ≥60 years) to estimate accuracy and compare to the bivariate model results.

To investigate heterogeneity, we generated forest plots of sensitivities and specificities for cutoff 10 for each study, first for all studies in each reference standard category, and then separately across participant subgroups within each reference standard category. We quantified cutoff 10 heterogeneity overall and across subgroups, by reporting estimated variances of the random effects for sensitivity and specificity (τ^2) and 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.[35] We used a complete case analysis since complete data for all subgrouping variables were available for 17,357 participants (98% of eligible participants in the database).

To estimate positive and negative predictive values using cutoff 10 for different major depression prevalence values, we generated nomograms for each reference standard category by applying the cutoff 10 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 for all items with at least 100 major depression cases among participants categorized as having “low” risk of bias and among participants with “high” or “unclear” risk of bias.

We did not conduct sensitivity analyses that combined IPDMA accuracy results with published results from studies that did not contribute IPD because among the 14 eligible studies that did not contribute IPD, only two studies with a semi-structured reference standard (N total = 173, N major depression = 29), one study with a fully structured reference standard (N total = 730, N MDD = 32), and one study using the MINI (N total = 172, N MDD = 33) published accuracy results eligible for the present IPDMA. The other studies had eligible datasets, but did not publish eligible diagnostic accuracy results (eTable1b).

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 quatrature point.

The only substantive deviations from our initial protocol were that we stratified accuracy results by reference standard category and did not conduct sensitivity analyses that combined IPDMA accuracy results with published results from studies that did not contribute IPD.

Patient and public involvement

Patients and members of the public were not involved in the study.

RESULTS

Search results and inclusion of primary datasets

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 (eFigure1). Reasons for exclusion for the 113 articles excluded at full-text level are provided in eTable1. In addition, authors of included studies contributed data from three unpublished studies, for a total of 58 datasets (N participants = 17,357, N major depression = 2,312 [13%]). Study characteristics of included studies and eligible studies that did not provide datasets are shown in eTable2a and eTable2b. Excluding the three unpublished studies, of 21,171 participants in 69 eligible published studies, 16,956 participants (80%) from 55 included published studies were included.

Of 58 included studies, 29 used semi-structured reference standards, 14 used fully structured reference standards, and 15 used the MINI (Table 1). The SCID was the most common semi-structured interview (26 studies, 4,733 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 232, 549, and 197, and mean number (%) with major depression were 32 (14%), 60 (11%), and 37 (19%; Table 2).

PHQ-9 accuracy by reference standard

Comparisons of sensitivity and specificity estimates by reference standard category are shown in Table 3. Cutoff 10 maximized combined sensitivity and specificity among studies using semi-structured interviews (sensitivity [95% CI] = 0.88 [0.83 to 0.92], specificity [95% CI] = 0.85 [0.82 to 0.88]). Based on cutoff 10, sensitivity and specificity [95% CI] were 0.70 [0.59 to 0.80] and 0.84 [0.77 to 0.89] for fully structured interviews, and 0.77 [0.68 to 0.83] and 0.87 [0.83 to 0.91]) for the

MINI. Across cutoffs, specificity estimates were similar across reference standards; however, sensitivity estimates for semi-structured interviews were 5 to 22% higher than for fully structured interviews (median difference = 18%, at cutoff 10) and 2 to 15% higher than for the MINI (median difference = 11%, at cutoff 10). ROC curves and AUC values are shown in eFigure2.

Heterogeneity analyses suggested moderate heterogeneity across studies, which improved in some instances when subgroups were considered. Cutoff 10 sensitivity and specificity forest plots are shown in eFigure3, with τ^2 and R values shown in eTable3.

Nomograms of positive and negative predictive values for cutoff 10 for each reference standard category are shown in Figure 1. For hypothetical major depression prevalence values of 5 to 25%, estimates of positive predictive values based on summary sensitivity and specificity values ranged from 24 to 66% for semi-structured interviews, 19 to 59% for fully structured interviews, and 24 to 66% for the MINI; estimates of negative predictive values ranged from 96 to 99% for semi-structured interviews, 89 to 98% for fully structured interviews, and 92 to 99% for the MINI.

When examined with meta-regression analysis, consistent with our main results, we found that PHQ-9 sensitivity estimates for semi-structured interviews were significantly higher than for fully structured interviews or the MINI (eTable4). The significant interactions corresponded to differences in sensitivity that across cutoffs were 4 to 22% higher for semi-structured interviews than for fully structured interviews (median = 18%) and 1 to 16% higher for semi-structured interviews than the MINI (median = 11%). Across all cutoffs, the magnitude of the differences estimated based on meta-regression were within 1% of those estimated using the original two-stage bivariate random effects meta-analytic models.

PHQ-9 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 eTable5 for results and eFigure4 for ROC curves and AUC values.

PHQ-9 accuracy among subgroups

For each reference standard category, comparisons of sensitivity and specificity estimates based on bivariate models across PHQ-9 cutoffs 5 to 15 among subgroups based on age, sex, country human development index and participant recruitment setting are shown in eTable5, with forest plots shown in eFigure3, ROC curves and AUC values shown in eFigure4, and τ^2 and R values shown in eTable3.

Of the total of 484 categorical subgroup analyses that were done (22 subgroups x 11 cutoff thresholds for sensitivity and specificity) using the bivariate model, 4 comparisons excluded the null value of zero difference for cutoffs 7 to 15. No comparisons that were significantly different in one reference standard category were statistically significant in either of the other two reference standard categories. Subgroup analyses are shown in eTable5.

In the meta-regression analyses, on the other hand, older age (measured continuously) was associated with higher specificity for all reference standards (eTable4). The significant interaction corresponded to specificity estimates that were 2 to 14% higher for participants aged ≥60 versus <60 among participants based on semi-structured interviews (median = 6%), 2 to 14% based on fully structured interviews (median = 8%), and 1 to 8% based on the MINI (median = 5%; eTable4). Across all cutoffs, the magnitudes of the differences estimated based on meta-regression with dichotomous age were within 2% of those estimated using the original two-stage bivariate random effects meta-analytic models.

Risk of bias sensitivity analyses

eTable6 shows QUADAS-2 ratings for each included primary study, while comparisons of PHQ-9 accuracy across individual items for each reference standard category are shown in eTable5. For the item on blinding of the reference standard to PHQ-9 results, specificity was significantly greater for studies and participants with high or unclear vs. low risk of bias for semi-structured interviews, but significantly greater for low vs. high or unclear risk of bias for fully structured interviews and the MINI. For the item on recruiting a consecutive or random sample of participants, specificity was significantly greater for low vs. high or unclear risk of bias for fully structured interviews and the MINI. No other statistically significant differences were found, and no significant differences replicated across all reference standards.

DISCUSSION

Principal findings

We compared the accuracy of scores on the PHQ-9 to detect major depression, separately, to semi-structured diagnostic interviews, fully structured diagnostic interviews (MINI excluded), and the MINI. Based on results from the semi-structured interviews, which most closely replicate clinical interviews done by trained professionals, the PHQ-9 was more sensitive than has been reported in previous meta-analyses that combined reference standards.[8,36] Specificity was similar to previous studies and across reference standards. Based on semi-structured interviews, the standard cutoff of 10 maximized combined sensitivity and specificity. There was evidence from multivariable meta-regression that the PHQ-9 may be more sensitive among older patients compared to younger patients, but this would not require that a different cutoff be used. Results did not differ depending on whether studies that did not explicitly exclude already diagnosed patients

were included or excluded. Among studies conducted in primary care settings, approximately half of patients who screen positive on the PHQ-9 had major depression.

Findings in context

This is the first meta-analysis that has analyzed diagnostic accuracy for the PHQ-9 separately for different diagnostic interviews. Diagnostic interviews that are used to classify major depression case status are imperfect reference standards. Semi-structured interviews, such as the SCID,[13] most closely approximate an expert diagnosis. They are set up to replicate a guided diagnostic conversation with standardized questions, but the option for interviewers to make additional queries and use clinical judgment to decide whether symptoms are present.[16,17] Semi-structured interviews involve lengthy processes that must be conducted by skilled diagnosticians and, thus, are expensive. Fully-structured interviews, such as the CIDI,[14] are designed to replicate as closely as possible expert-administered semi-structured interviews, but are not expected to have the same level of validity and reliability. Fully structured interviews can be administered by lay interviewers and involve fully scripted standardized interview protocols that are read verbatim without additional probes or interpretation. Fully structured interviews are designed to increase reliability with administration by lay interviewers who are are not trained to independently carry out diagnostic interviews at the possible cost of validity.[16,17] The MINI is a specific fully structured interview that was designed to be administered in a fraction of the time compared to other interviews and described by its developers as intentionally over-inclusive.[19,20] Test-retest reliability for diagnosis of current major depression has been reported to be kappa = 0.74 for the SCID (N = 51; mean = 9 days)[37] and kappa = 0.52 for the CIDI (N = 60, mean = 2 days).[38]

Consistent with the design features and rigour of each type of diagnostic interview, we previously reported that compared to semi-structured interviews, fully structured interviews

(excluding the MINI) classify more people with low symptoms as having major depression but fewer people with high symptoms.[12] We also found that the MINI identified approximately twice as many cases as other fully structured interviews.[12] The finding in the present study that sensitivity was greater among studies with semi-structured rather than fully structured reference standards is consistent with both the design features and rigor of the different types of diagnostic interviews and with our previous findings. It is possible that the lower sensitivity among fully structured interviews may have been due to overdiagnosis of major depression among participants with low depressive symptom levels when fully structured interviews were used. In the present meta-analysis, most participants did not have major depression (87%), thus misclassification of major depression among participants with sub-threshold depressive symptom levels based on fully structured interviews might explain the lower sensitivity compared to semi-structured interviews if the PHQ-9 were less likely to identify "false positive" classifications based on fully structured interviews. The same logic would apply to the lower sensitivity for the MINI.

Among studies that used semi-structured reference standards, sensitivity was also greater than reported in previous traditional meta-analyses, where studies with semi- and fully structured reference standards and the MINI were combined without adjustment. Using IPD data from the 29 studies that used a semi-structured interview as the reference standard, we found that at cutoff 10, sensitivity and specificity were 0.88 and 0.85 compared to 0.78 and 0.87 in a 2015 conventional meta-analysis of 34 studies that combined reference standards.[8] In primary care settings, we found sensitivity and specificity of 0.94 and 0.88 (9 studies with a semi-structured interview) compared to 0.82 and 0.85 in a 2016 conventional meta-analysis of 20 studies that combined reference standards.[36]

For semi-structured interviews, major depression prevalence in our dataset was 14%. Using our cutoff 10 accuracy estimates (sensitivity = 0.88, specificity = 0.85), positive predictive value would only be 49%; thus 51% of all positive screens would be false positives. For primary care settings, where accuracy was even higher, major depression prevalence was 12%. Using our accuracy estimates for cutoff 10 (sensitivity = 0.94, specificity = 0.88, positive predictive value = 52%), 22% of patients in primary care would screen positive at this cutoff, but only approximately half would be true positives.

Clinical implications

Screening for depression in primary care is recommended in the United States,[39] but national guidelines from Canada and the United Kingdom recommend against routine depression screening.[40,41] Those guidelines cite the lack of evidence of benefit from well-conducted randomised controlled trials, as well as concerns about high false positive rates, overdiagnosis, and substantial resource utilization and opportunity costs.[40-41] Well-conducted and adequately powered trials designed specifically to assess the effects of depression screening are needed.[1,2,40-43] If screening is to be done clinically based on recommendations in the United States, the cutoff that maximizes sensitivity and specificity is the standard cutoff of 10 or greater. It is not known, however, if using this standard cutoff would maximize the likelihood that screening would successfully improve mental health and minimize unnecessary resource use and adverse outcomes if tested in a trial. Ideally, robust trials that are sufficiently powered to evaluate the effects of screening across a range of cutoffs will be conducted. Clinical trials provide the best possible evidence to inform both the decision on whether or not depression screening should be implemented as part of routine care and, if so, on thresholds for intervening or what steps might be taken for patients with borderline screening results.[44]

Strengths and limitations

This was the first study to use IPDMA to assess diagnostic accuracy of the PHQ-9 or any other depression screening tool. Strengths include the large sample size, the ability to include results from all cutoffs from all studies (rather than just those published), the ability to examine participant subgroups, and the ability to assess accuracy separately across reference standards, which had not been done previously. There are also limitations to consider. First, we were unable to include primary data from 14 of 69 published eligible datasets (20% of eligible datasets and participants), and we restricted our analyses to those with complete data for all variables used in our various analyses (98% of available data). Nonetheless, for all cutoffs other than 10, our sample was much larger than previous traditional meta-analyses of the PHQ-9. Second, despite the large sample size, there was substantial heterogeneity across studies, although it did improve in some instances when subgroups were considered. We were not able to conduct subgroup analyses based on specific medical comorbidities or cultural aspects such as country or language because comorbidity data were not available for over half of participants, and many countries and languages were represented in few primary studies. However, we were able to compare participant subgroups based on age, sex, country human development index, and participant recruitment setting category, which has not been done previously. Third, while we categorized studies based on the diagnostic interview administered, interviews are sometimes adapted and thus not always 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.

Although our original two-stage bivariate random effects meta-analytic models did not find significant differences in accuracy estimates across participant subgroups, our meta-regressions

suggested that specificity might be somewhat higher among older participants whether measured continuously or dichotomously. This difference in significance may be due to the differences between the analytical approaches. Whereas statistical significance of the interactions between covariates and accuracy estimates in the meta-regressions were based on parametric standard errors, statistical significance of subgroup comparisons in the two-stage bivariate random effects models was based on non-parametric bootstrap methods. Moreover, whereas the meta-regression models provide a within-study interpretation, the two-stage bivariate random effects models did not link study clusters across subgroups and thus focused more on between-study comparisons.

Conclusions and policy implications

In summary, we found that PHQ-9 sensitivity compared to semi-structured reference standards was substantially greater than when compared to fully structured reference standards or the MINI. It was also substantially higher than previously reported in conventional meta-analyses which combined reference standards.[8,36] The standard cutoff of 10 or greater maximized combined sensitivity and specificity. However, in primary care, approximately half of patients with positive screens would be false positives if used in practice, a concern that has been emphasized by the Canadian Task Force on Preventive Health Care, UK National Screening Committee, and UK National Institute for Health and Care Excellence, given the resources that would be required for additional assessment and the possibility that some of these patients might be treated without benefit.[40,41,43] Future research on the PHQ-9 should ideally be based on semi-structured diagnostic interviews, should consider estimating probabilities of depression across the full spectrum of PHQ-9 screening scores (rather than dichotomizing scores at a cutoff), and should combine screening scores with individual characteristics to generate individualized probabilities of major depression.

Contributions:

BLevis, AB, JB, PC, SG, JPAI, LAK, DM, SBP, IS, RCZ and BDT were responsible for the study conception and design. JB and LAK designed and conducted database searches to identify eligible studies. DHA, BA, LA, HRB, MB, CHB, PB, GC, MHC, JCNC, KC, YC, JMG, JD, JRF, FHF, DF, BG, FGS, CGG, BJH, JH, PAH, MHärter, UH, LH, SEH, MHudson, MI, KI, NJ, MEK, KMK, YK, SL, ML, SRL, BLöwe, LM, AM, SMS, TNM, KM, FLO, VP, BWP, PP, AP, KR, AGR, ISS, JS, ASidebottom, ASimning, LS, SCS, PLLT, AT, CMvdFC, HCvW, PAV, JW, MAH, KW, MY, YZ, and BDT contributed primary datasets that were included in this study. BLevis, KER, NS, MA, DBR, MJC, TAS, and BDT contributed to data extraction and coding for the meta-analysis. BLevis, AB, AWL, and BDT contributed to the data analysis and interpretation. BLevis, AB, and BDT contributed to drafting the manuscript. All authors provided a critical review and approved the final manuscript. AB and BDT are the guarantors; they had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses. The corresponding author (BT) attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

Funding:

This study was funded by the Canadian Institutes of Health Research (CIHR; KRS-134297). Ms. Levis was supported by a CIHR Frederick Banting and Charles Best Canada Graduate Scholarship doctoral award. Drs. Benedetti and Thombs were supported by Fonds de recherche du Québec - Santé (FRQS) researcher salary awards. Ms. Riehm and Ms. Saadat were supported by CIHR Frederick Banting and Charles Best Canada Graduate Scholarship master's awards. Mr. Levis and Ms. Azar were supported by FRQS Masters Training Awards. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. Collection of data for the study by Arroll et al. was supported by a project grant from the Health Research Council of New Zealand. Data collection for the study by Ayalon et al. was supported from a grant from Lundbeck International. The primary study by Khamseh et al. was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary study by Bombardier et al. was supported by the Department of Education, National Institute on Disability and Rehabilitation Research, Spinal Cord Injury Model Systems: University

of Washington (grant no. H133N060033), Baylor College of Medicine (grant no. H133N060003), and University of Michigan (grant no. H133N060032). Dr. Butterworth was supported by Australian Research Council Future Fellowship FT130101444. 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. Dr. Conwell received support from NIMH (R24MH071604) and the Centers for Disease Control and Prevention (R49 CE002093). Collection of data for the primary study by Delgadillo et al. was supported by grant from St. Anne's Community Services, Leeds, United Kingdom. Collection of data for the primary study by Fann et al. was supported by grant RO1 HD39415 from the US National Center for Medical Rehabilitation Research. The primary 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 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). 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. 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. The study by van Steenbergen-Weijenburg et al. was funded by Innovatiefonds Zorgverzekeraars. Dr. Vöhringer was supported by the Fund for Innovation and Competitiveness of the Chilean Ministry of Economy, Development and Tourism, through the Millennium Scientific

Initiative (grant number IS130005). 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. 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. No other authors reported funding for primary studies or for their work on the present study.

Declaration of Competing Interests:

All authors have completed the ICJME uniform disclosure form at www.icmje.org/coi_disclosure.pdf 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.

Ethical Approval: As this study involved secondary analysis of anonymized previously collected data, the Research Ethics Committee of the Jewish General Hospital declared that this project did not require research ethics approval. However, for each included dataset, we confirmed that the original study received ethics approval and that all patients provided informed consent.

Transparency Declaration: The manuscript's guarantor affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Data Sharing: Requests to access data should be made to the corresponding author at brett.thombs@mcgill.ca.

What is already known on this topic:

- The PHQ-9 is the most commonly used depression screening tool in primary care.
- Previous meta-analyses on the diagnostic test accuracy of the PHQ-9 have been limited by selective cutoff reporting in primary studies; the inability to assess differences across patient subgroups; the inability to exclude participants already diagnosed or being treated for depression, who would not be screened in practice; and the combining of accuracy estimates without differentiating between reference standards.

What this study adds:

- PHQ-9 diagnostic accuracy when compared to diagnoses made by semi-structured diagnostic interviews is greater than when compared to diagnoses made by other reference standards and greater than reported in previous meta-analyses, which did not distinguish between different diagnostic standards.
- PHQ-9 diagnostic accuracy does not differ substantively across participant subgroups except for age, where it may be more specific among older patients.
- The standard cutoff of 10 or greater maximizes combined sensitivity and specificity overall and for subgroups.

PRINT ABSTRACT

Study question: What is the diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression?

Methods: Individual participant data meta-analysis was used to synthesize results from studies that compared PHQ-9 scores to major depression diagnoses from validated diagnostic interviews. For PHQ-9 cutoffs 5 to 15, bivariate random-effects meta-analysis was used to estimate pooled sensitivity and specificity among studies that used semi-structured diagnostic interviews, fully structured interviews, and the Mini International Neuropsychiatric (MINI), separately. Sensitivity and specificity were examined among participant subgroups and, separately, using meta-regression, considering all subgroup variables in a single model.

Study answer and limitations: Data were obtained for 58 of 72 eligible studies (N participants = 17,357, N cases = 2,312). Combined sensitivity and specificity was maximized at a cutoff of ≥10 among studies using a semi-structured interview; sensitivity [95% CI] was 0.88 [0.83 to 0.92], specificity [95% CI] was 0.85 [0.82 to 0.88]). Across cutoffs, sensitivity with semi-structured interviews was higher than for fully structured interviews (MINI excluded) and for the MINI. Specificity was similar across diagnostic interviews. In studies conducted in primary care using a semi-structured interview (major depression prevalence = 12%), sensitivity [95% CI] was 0.94 [0.88 to 0.97] and specificity [95% CI] was 0.88 [0.79 to 0.93]). Study limitations include the inability to obtain data for 14 eligible studies, substantial heterogeneity across included studies, and the inability to conduct subgroup analyses based on specific medical comorbidities or cultural aspects.

What this study adds: PHQ-9 sensitivity compared to semi-structured diagnostic interviews was greater than in previous meta-analyses that combined reference standards. A cutoff of ≥10 maximized combined sensitivity and specificity overall and for subgroups.

Funding: The Canadian Institutes of Health Research (KRS-134297; PCG-155468).

Registration: PROSPERO (CRD42014010673).

REFERENCES

1. Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care? *BMJ*. 2014;**348**:g1253.
2. Thombs BD, Coyne JC, Cuijpers P, et al. Rethinking recommendations for screening for depression in primary care. *CMAJ* .2012;184:413-8.
3. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;**16**:606-13.
4. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann.* 2002;**32**:1-7.
5. 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**:1737-44.
6. Wittkampf KA, Naeije L, Schene AH, et al. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry.* 2007;**29**:388-95.
7. Gilbody S, Richards D, Brealey S, et al. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med.* 2007;**22**:1596-1602.
8. 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**:567-76.
9. Levis B, Benedetti A, Levis AW, et al. Selective cutoff reporting in studies of diagnostic test accuracy: a comparison of conventional and individual-patient-data meta-analysis of the Patient Health Questionnaire-9 depression screening tool. *Am J Epidemiol*. 2017;**185**:954-64.

10. Thombs BD, Arthurs E, El-Baalbaki G, et al. 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.
11. 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:e0150067.
12. 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**:377-85.
13. First MB. Structured clinical interview for the DSM (SCID). John Wiley & Sons, Inc. 1995.
14. 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**:1069-77.
15. 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**:1001-13.
16. 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**:1013-20.
17. Nosen E, Woody SR. Chapter 8: Diagnostic Assessment in Research. In, McKay D. Handbook of research methods in abnormal and clinical psychology. Sage; 2008.
18. Kurdyak PA, Gnam WH. Small signal, big noise: performance of the CIDI depression module. *Can J Psychiatry*. 2005;**50**:851-6.

19. 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**:224-31.
20. 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**:232-41.
21. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;**340**:c221.
22. 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:**27**;3:124.
23. 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):388-96.
24. Stewart LA, Clarke M, Rovers M, et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA*. 2015;**313**(16):1657-65.
25. PRESS – Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E). Ottawa: CADTH; 2016.
26. United Nations. International Human Development Indicators. http://hdr.undp.org/en/countries. Accessed February 1, 2019.

27. 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**:529-36.
28. World Health Organization. Schedules for clinical assessment in neuropsychiatry: manual. Amer Psychiatric Pub Inc. 1994.
29. 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**:897-905.
30. 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**:465-86.
31. 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**:381-9.
32. Riley RD, Dodd SR, Craig JV, et al. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med.* 2008;**27**:6111-36.
33. 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.
34. 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; 2008:401-33.
35. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* 2002;**21**:1539-58.

36. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. *BJPsych Open*. 2016;**2**:127-38.
37. Shankman SA, Funkhouser CJ, Klein DN, et al. Reliability and validity of severity dimensions of psychopathology assessed using the Structured Clinical Interview for DSM-5 (SCID). *Int J Methods Psychiatr Res*. 2018;**27**:e1590.
38. Semler G, Wittchen HU, Joschke K, et al. Test-retest reliability of a standardized psychiatric interview (DIS/CIDI). *Eur Arch Psychiatry Neurol Sci*. 1987;**236**:214-22
39. 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**:380-7.
40. Allaby M. Screening for depression: A report for the UK National Screening Committee (Revised report). London, United Kingdom: UK National Screening Committee; 2010.
41. Joffres M, Jaramillo A, Dickinson J, et al. Recommendations on screening for depression in adults. *CMAJ*. 2013;**185**:775-82.
42. 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.
43. National Institute for Health and Care Excellence. Depression in Adults: treatment and management. Consultation draft (May 2018). https://www.nice.org.uk/guidance/gid-cgwave0725/documents/full-guideline-updated. Accessed February 1, 2019.
44. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ*. 2012;**344**:e686.

FIGURES

Figure 1. Nomograms of positive and negative predictive value for cutoff 10 of the PHQ-9 for each reference standard category

Nomograms of a) positive predictive value and b) negative predictive value for cutoff 10 of the PHQ-9, for major depression prevalence values of 5 to 25%, for semi-structured diagnostic interviews, fully structured diagnostic interviews, and the MINI.

TABLES

Table 1. Participant data by diagnostic interview

Diagnostic Interview	N Studies	N Participants	N (%) Major Depression
Semi-structured			
SCID	26	4,733	785 (17)
SCAN	2	1,892	130 (7)
DISH	1	100	9 (9)
Fully structured			
CIDI	11	6,272	554 (9)
DIS	1	1,006	221 (22)
CIS-R	2	402	64 (16)
MINI	15	2,952	549 (19)
Total	58	17,357	2,312 (13)

Abbreviations: 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

Table 2. Participant data by subgroup

Participant Subgroup	Semi-Structured Diagnostic Interviews			Fully Structured Diagnostic Interviews			MINI		
	N Studies	N Participants	N (%) Major Depression	N Studies	N Participants	N (%) Major Depression	N Studies	N Participants	N (%) Major Depression
All participants	29	6,725	924 (14)	14	7,680	839 (11)	15	2,952	549 (19)
Participants not currently diagnosed or receiving treatment for a mental health problem	20	2,942	421 (14)	6	4,161	306 (7)	6	927	168 (18)
Age <60	26	4,132	629 (15)	14	5,504	645 (12)	14	1,958	310 (16)
Age ≥60	24	2,577	295 (11)	10	2,175	194 (9)	13	979	239 (24)
Women	28	3,906	573 (15)	14	4,285	463 (11)	15	1,666	337 (20)
Men	25	2,812	351 (12)	13	3,395	376 (11)	15	1,286	212 (16)
Very high country human development index	25	6,195	739 (12)	9	5,740	592 (10)	10	1,924	430 (22)
High country human development index	4	530	185 (35)	2	326	61 (19)	3	542	61 (11)
Low-medium country human development index	--	--	--	3	1,614	186 (12)	2	486	58 (12)
Non-medical care	2	567	105 (19)	2	963	74 (8)	2	299	72 (24)
Primary care	9	3,163	377 (12)	5	3,578	273 (8)	5	1,290	168 (13)
Inpatient specialty care	8	867	121 (14)	2	372	34 (9)	1	137	25 (18)
Outpatient specialty care	12	2,128	321 (15)	5	2,767	458 (17)	7	1,226	284 (23)

[a]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 total number in the reference category

Table 3a. Comparison of sensitivity and specificity estimates among semi-structured vs. fully structured reference standards

	Semi-Structured Reference Standard[a]		Fully Structured Reference Standard[b]		Difference across reference standards (Semi-structured - Fully structured)[c]	
Cutoff	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**
5	0.98 (0.96 to 0.99)	0.55 (0.49 to 0.60)	0.93 (0.87 to 0.97)	0.54 (0.43 to 0.64)	0.05 (-0.01 to 0.13)	0.01 (-0.13 to 0.16)
6	0.98 (0.95 to 0.99)	0.63 (0.58 to 0.67)	0.91 (0.83 to 0.95)	0.61 (0.51 to 0.71)	0.07 (-0.01 to 0.18)	0.02 (-0.12 to 0.17)
7	0.98 (0.94 to 0.99)	0.69 (0.65 to 0.74)	0.86 (0.75 to 0.92)	0.69 (0.59 to 0.77)	0.12 (0.00 to 0.26)	0.00 (-0.10 to 0.15)
8	0.95 (0.91 to 0.97)	0.75 (0.71 to 0.79)	0.82 (0.71 to 0.89)	0.75 (0.66 to 0.82)	0.13 (0.00 to 0.28)	0.00 (-0.10 to 0.13)
9	0.91 (0.87 to 0.94)	0.80 (0.77 to 0.83)	0.74 (0.63 to 0.83)	0.79 (0.72 to 0.86)	0.17 (0.05 to 0.34)	0.01 (-0.08 to 0.12)
10	0.88 (0.83 to 0.92)	0.85 (0.82 to 0.88)	0.70 (0.59 to 0.80)	0.84 (0.77 to 0.89)	0.18 (0.04 to 0.36)	0.01 (-0.05 to 0.12)
11	0.84 (0.78 to 0.89)	0.89 (0.86 to 0.91)	0.62 (0.51 to 0.72)	0.87 (0.81 to 0.91)	0.22 (0.07 to 0.40)	0.02 (-0.04 to 0.10)
12	0.79 (0.73 to 0.83)	0.91 (0.89 to 0.93)	0.57 (0.45 to 0.68)	0.89 (0.85 to 0.93)	0.22 (0.05 to 0.40)	0.02 (-0.03 to 0.09)
13	0.70 (0.65 to 0.75)	0.93 (0.91 to 0.95)	0.49 (0.38 to 0.61)	0.92 (0.89 to 0.95)	0.21 (0.04 to 0.40)	0.01 (-0.03 to 0.07)
14	0.64 (0.58 to 0.70)	0.95 (0.93 to 0.96)	0.44 (0.32 to 0.56)	0.94 (0.91 to 0.96)	0.20 (0.03 to 0.40)	0.01 (-0.02 to 0.05)
15	0.56 (0.50 to 0.62)	0.96 (0.95 to 0.97)	0.35 (0.25 to 0.46)	0.96 (0.93 to 0.97)	0.21 (0.05 to 0.39)	0.00 (-0.02 to 0.04)

[a] N Studies = 29; N Participants = 6,725; N major depression = 924

[b] N Studies = 14; N Participants = 7,680; N major depression = 839

[c] 1 bootstrap iteration (0.01%) did not produce a difference estimate for cutoff 5. This iteration was removed prior to determining the bootstrapped CI.

Abbreviations: CI: confidence interval

Table 3b. Comparison of sensitivity and specificity estimates among semi-structured vs. MINI reference standards

	Semi-Structured Reference Standard[a]		MINI Reference Standard[b]		Difference across reference standards (Semi-structured - MINI)	
Cutoff	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**
5	0.98 (0.96 to 0.99)	0.55 (0.49 to 0.60)	0.96 (0.93 to 0.98)	0.57 (0.50 to 0.64)	0.02 (-0.02 to 0.07)	-0.02 (-0.14 to 0.11)
6	0.98 (0.95 to 0.99)	0.63 (0.58 to 0.67)	0.93 (0.87 to 0.97)	0.66 (0.59 to 0.72)	0.05 (-0.01 to 0.12)	-0.03 (-0.13 to 0.09)
7	0.98 (0.94 to 0.99)	0.69 (0.65 to 0.74)	0.90 (0.82 to 0.94)	0.72 (0.66 to 0.78)	0.08 (-0.00 to 0.16)	-0.03 (-0.12 to 0.08)
8	0.95 (0.91 to 0.97)	0.75 (0.71 to 0.79)	0.86 (0.78 to 0.91)	0.78 (0.73 to 0.83)	0.09 (-0.01 to 0.19)	-0.03 (-0.11 to 0.06)
9	0.91 (0.87 to 0.94)	0.80 (0.77 to 0.83)	0.82 (0.72 to 0.88)	0.84 (0.79 to 0.87)	0.09 (-0.02 to 0.22)	-0.04 (-0.09 to 0.05)
10	0.88 (0.83 to 0.92)	0.85 (0.82 to 0.88)	0.77 (0.68 to 0.83)	0.87 (0.83 to 0.90)	0.11 (-0.01 to 0.25)	-0.02 (-0.07 to 0.06)
11	0.84 (0.78 to 0.89)	0.89 (0.86 to 0.91)	0.70 (0.62 to 0.77)	0.90 (0.86 to 0.92)	0.14 (0.01 to 0.30)	-0.01 (-0.06 to 0.05)
12	0.79 (0.73 to 0.83)	0.91 (0.89 to 0.93)	0.65 (0.56 to 0.72)	0.92 (0.89 to 0.94)	0.14 (-0.01 to 0.28)	-0.01 (-0.05 to 0.05)
13	0.70 (0.65 to 0.75)	0.93 (0.91 to 0.95)	0.57 (0.49 to 0.65)	0.94 (0.91 to 0.96)	0.13 (-0.03 to 0.26)	-0.01 (-0.04 to 0.04)
14[c]	0.64 (0.58 to 0.70)	0.95 (0.93 to 0.96)	0.49 (0.42 to 0.56)	0.96 (0.93 to 0.97)	0.15 (0.01 to 0.28)	-0.01 (-0.04 to 0.03)
15[c]	0.56 (0.50 to 0.62)	0.96 (0.95 to 0.97)	0.42 (0.35 to 0.49)	0.97 (0.95 to 0.98)	0.14 (-0.01 to 0.27)	-0.01 (-0.03 to 0.02)

[a] N Studies = 29; N Participants = 6,725; N major depression = 924

[b] N Studies = 15; N Participants = 2,952; N major depression = 549

[c] For these cutoffs, among studies that used the MINI as the reference standard, the default optimizer in glmer failed, thus bobyqa was used instead.

Abbreviations: CI: confidence interval; MINI: Mini International Neuropsychiatric Interview