Supplementary Material

Affiliations of DEPRESSD Collaboration members

Supplementary methods A. Search strategies

Supplementary methods B. QUADAS-2 Coding manual for primary studies included in the present study

Supplementary figure A. Flow diagram of study selection process

Supplementary figure B. ROC curves for each reference standard category

Supplementary figure C. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 for each reference standard category, including participant subgroups based on age, sex, human development index and care setting (Note that some confidence intervals are very wide due to small numbers of cases/non-cases in certain subgroups)

Supplementary figure D. ROC curves of subgroups for each reference standard category

Supplementarty table A. Reasons for exclusion for all articles excluded at full-text level

Supplemetary table B. Characteristics of included primary studies as well as eligible primary studies not included in the present study

Supplementary table C. Estimates of heterogeneity at PHQ-9 cutoff score of 10

Supplementary table D. Model coefficients and p-values from post-hoc one-stage meta-regressions

Supplementary table E. Comparison of PHQ-9 sensitivity and specificity estimates among participants not currently diagnosed or receiving treatment for a mental health problem compared to all participants as well as among participant subgroups based on age, sex,

human development index, care setting, and risk of bias factors, for each reference standard category

Supplementary table F. QUADAS-2 ratings for each primary study included in the present study

Affiliations of DEPRESSD Collaboration members

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)

Supplementary methods A. Search strategies

MEDLINE (OvidSP)

1. PHQ*.af.
2. patient health questionnaire*.af.
3. 1 or 2
4. Mass Screening/
5. Psychiatric Status Rating Scales/
6. "Predictive Value of Tests"/
7. "Reproducibility of Results"/
8. exp "Sensitivity and Specificity"/
9. Psychometrics/
10. Prevalence/
11. Reference Values/
12.. Reference Standards/
13. exp Diagnostic Errors/
14. Mental Disorders/di, pc [Diagnosis, Prevention & Control]
15. Mood Disorders/di, pc [Diagnosis, Prevention & Control]
16. Depressive Disorder/di, pc [Diagnosis, Prevention & Control]
17. Depressive Disorder, Major/di, pc [Diagnosis, Prevention & Control]
18. Depression, Postpartum/di, pc [Diagnosis, Prevention & Control]
19. Depression/di, pc [Diagnosis, Prevention & Control]
20. validation studies.pt.
21. comparative study.pt.
22. screen*.af.
23. prevalence.af.
24. predictive value*.af.
25. detect*.ti.
26. sensitiv*.ti.
27. valid*.ti.
28. revalid*.ti.
29. predict*.ti.
30. accura*.ti.
31. psychometric*.ti.
32. identif*.ti.
33. specificit*.ab.
34. cut?off*.ab.
35. cut* score*.ab.
36. cut?point*.ab.
37. threshold score*.ab.
38. reference standard*.ab.
39. reference test*.ab.
40. index test*.ab.
41. gold standard.ab.
42. or/4-41
43. 3 and 42

44. limit 43 to yr=”2000-Current”

PsycINFO (OvidSP)

1. PHQ*.af.
2. patient health questionnaire*.af.
3. 1 or 2
4. Diagnosis/
5. Medical Diagnosis/
6. Psychodiagnosis/
7. Misdiagnosis/
8. Screening/
9. Health Screening/
10. Screening Tests/
11. Prediction/
12. Cutting Scores/
13. Psychometrics/
14. Test Validity/
15. screen*.af.
16. predictive value*.af.
17. detect*.ti.
18. sensitiv*.ti.
19. valid*.ti.
20. revalid*.ti.
21. accura*.ti.
22. psychometric*.ti.
23. specificit*.ab.
24. cut?off*.ab.
25. cut* score*.ab.
26. cut?point*.ab.
27. threshold score*.ab.
28. reference standard*.ab.
29. reference test*.ab.
30. index test*.ab.
31. gold standard.ab.
32. or/4-31
33. 3 and 32
38. Limit 33 to “2000 to current”

Web of Science (Web of Knowledge)

#1: TS=(PHQ* OR “Patient Health Questionnaire*”)

#2: TS= (screen* OR prevalence OR “predictive value*” OR detect* OR sensitiv* OR valid* OR revalid* OR predict* OR accura* OR psychometric* OR identif* OR specificit* OR cutoff* OR “cut off*” OR “cut* score*” OR cutpoint* OR “cut point*” OR “threshold score*” OR “reference standard*” OR “reference test*” OR “index test*” OR “gold standard”)
#1 AND #2
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH Timespan=2000-2014

Supplementary methods B. QUADAS-2 Coding manual for primary studies included in the present study

Domain 1: Participant Selection

1. **Signalling question 1 – Was a consecutive or random sample of patients enrolled?:** Code as "yes" if a consecutive or random sample of participants were recruited for the study and the percentage of eligible participants who participate is ≥75%. If the study indicates that consecutive or random participants were recruited, but does not give an indication of the total number of eligible participants and how many agreed to participate in the study, this should be rated "unclear". If the percentage of eligible participants included in the study was between ≥50% and <75%, then this should also be marked as "unclear". If a very low rate of eligible participants (<50%) were included in the study, this should be coded "no." In "Notes", please provide the relevant numbers and percentages used to make a determination. If a convenience sample of participants was recruited for the study or if the study was a case-control design, code as "no".

2. **Signalling question 2 – Was a case-control design avoided?:** Code as "yes" if the study did not employ a case-control design. Code as "no" if the study used a case-control design.

3. **Signalling question 3 – Did the study avoid inappropriate exclusions?:** Inappropriate exclusions refer to situations where an important part of the screening population was excluded from the study based on characteristics that could be related to screening results. Code as "yes" if the study does not inappropriately exclude participants. Code as "no" if the study inappropriately excludes participants.

4. **Overall risk of bias**: Rate as "low", "high", or "unclear" as described in QUADAS-2. Please indicate factors in decision in "Notes". NOTE: if signalling question 1 was coded "Unclear" the overall risk of bias is either a) Unclear, in cases where the denominator is not specified, or the percentage cannot be calculated, or method of participant selection is unclear OR b) Low, in cases where the percentage can be calculated, and is between 50-75%. If signalling question 1 is a "no" and signalling questions 2 and 3 are both "yes" then the risk of bias is coded "Unclear".

5. **Applicability concerns:** Code as "low" if study excluded participants who were already diagnosed or treated for depression or if the study included these patients, but they can be excluded using the individual patient data. Also code as "low" if the study did not exclude participants already diagnosed with depression and the overall percentage of these participants is low (e.g., ≤ 2.0% of total participants), even if there is not a variable to exclude them. Code "unclear" if the study did not exclude participants already diagnosed or treated for depression and it is not known how many diagnosed and treated patients were included or if the percentage is moderate (e.g., >2.0% but ≤ 5.0%). Code "high" if already diagnosed and treated patients are included and make up > 5.0% of the total sample and there is not a variable to exclude them. Please see aggregated study information sheet to code this.

Domain 2: Index Test

1. **Signalling question 1 - Were the index test results interpreted without the knowledge of the results of the reference standard?:** Code this item as "N/A" for all studies, as the index test is scored and does not require interpretation.

2. **Signalling question 2 - If a threshold was used, was it pre-specified?:** Code this item as "N/A" for all studies, as individual participant data allows for testing at all thresholds/cut-offs.

3. **Overall risk of bias:** Rate this item as "low" for all studies since the interpretation of the index test is fully automated in scoring self-report depressive symptom questionnaires and the individual participant data allows for testing at all thresholds/cut-offs.

4. **Applicability concerns:** Code "low" if the standard language version of the index test was used or if a translated version was used with an appropriate translation and back-translation process, or a translated version is located online. Code "unclear" if a translated version was used and it is not clear what steps were taken to ensure the quality of the translation or if only forward translation was used.

Domain 3: Reference Standard

1. **Signalling question 1 – Is the reference standard likely to correctly classify the condition?:** This question will be coded as "yes" for all studies because the use of a validated semi- or fully-structured psychiatric interview to assess participants for a DSM or ICD diagnosis of MDD/MDE is an eligibility requirement.

2. **Signalling question 2 – Were the reference standard results interpreted without knowledge of the results of the index test?:** Code as "yes" if the person administering the diagnostic interview was blinded to the participant's score on the index test, or if the diagnostic interview was administered before the index test. Code as "no" if the person administering the diagnostic interview was not blinded or was aware of the participant's score on the index test. Code as "unclear" if the study does not indicate whether blinding occurred and we cannot ascertain whether blinding occurred.

3. **Study-specific Signalling question 3 – Did a qualified person administer the reference standard?:** For structured clinical interviews, this will typically be coded "yes" as no specific clinical training is required. For semi-structured interviews, this will be coded "yes" if a trained diagnostician administered the clinical interview (e.g., psychiatrist, psychologist, social worker). Code "no" if individuals without the required training administered the reference standard (e.g., students, research assistants). Code "unclear" if the characteristics of personnel who administered the diagnostic interview cannot be ascertained or if advanced trainees, such as doctoral students, administered the reference standard. If the name of the interviewer is provided in the article, but no credentials are listed, then code based on credentials retrieved online for the interviewer.

4. **Overall risk of bias:** The coding of this item should consider blinding of the person administering the diagnostic interview to the participant's score on the index test and the qualifications of individuals administering the reference standard interview.

5. **Applicability concerns:** This item will be coded as "low" for most standard language studies, since the use of a validated semi- or fully-structured psychiatric interview to assess participants for a DSM or ICD diagnosis of MDD/MDE is an eligibility requirement. For translated versions of a validated reference standard, code "low" if a translated version was used with an appropriate translation and back-translation process, or a translated version is located online. Code "unclear" if a translated version was used and it is not clear what steps were taken to ensure the quality of the translation or if only forward translation was used.

Domain 4: Flow and Timing

1. **Signalling question 1 – Was there an appropriate interval between index test and reference standard?:** Only patient data with two weeks or less between the index text and reference standard are included. Thus, code "yes" if index test and reference standard were administered within a week of each other. Code "unclear" if the period was greater than one week (but less than two weeks) or if the timing cannot be ascertained beyond knowing that it was < 2 weeks. Note that this item may be coded differently for different patients from the same study. Please see aggregated study information sheet to code this.

2. **Signalling question 2 – Did all patients receive a reference standard?:** This will typically be coded "yes". If a portion of positive and negative screens receive the reference standard, and the patients selected were chosen randomly, code "yes". If non-random selection based on clinical factors or the index test determined whether or not patients received a reference standard, then code "unclear" or "no". An example of all patients not receiving a reference standard would occur, for instance, if patients who

endorsed suicidality on the index test were referred for evaluation and did not receive the reference standard interview.

3. **Signalling question 3 – Did all patients receive the same reference standard?:** This question will typically be coded as "yes" for all studies, since the reference standard is almost always consistent within each study.

4. **Signalling question 4 – Were all patients included in the analysis?:** When coding for this question, compare the number of participants who received the index test to the number of participants who received the reference standard. Code as "yes" if at least 90% of participants who received the index test also received the reference standard, or vice versa, and were included in analyses. Code as "unclear" if this difference is ≥ 80%, but < 90% or if it cannot be determined. Code as "no" if it is < 80%. If the study used randomly selected patients for either the index test or the reference standard, do not count the participants who did not receive the reference standard for that reason as missing. In "Notes", please provide the relevant numbers and percentages used to make a determination.

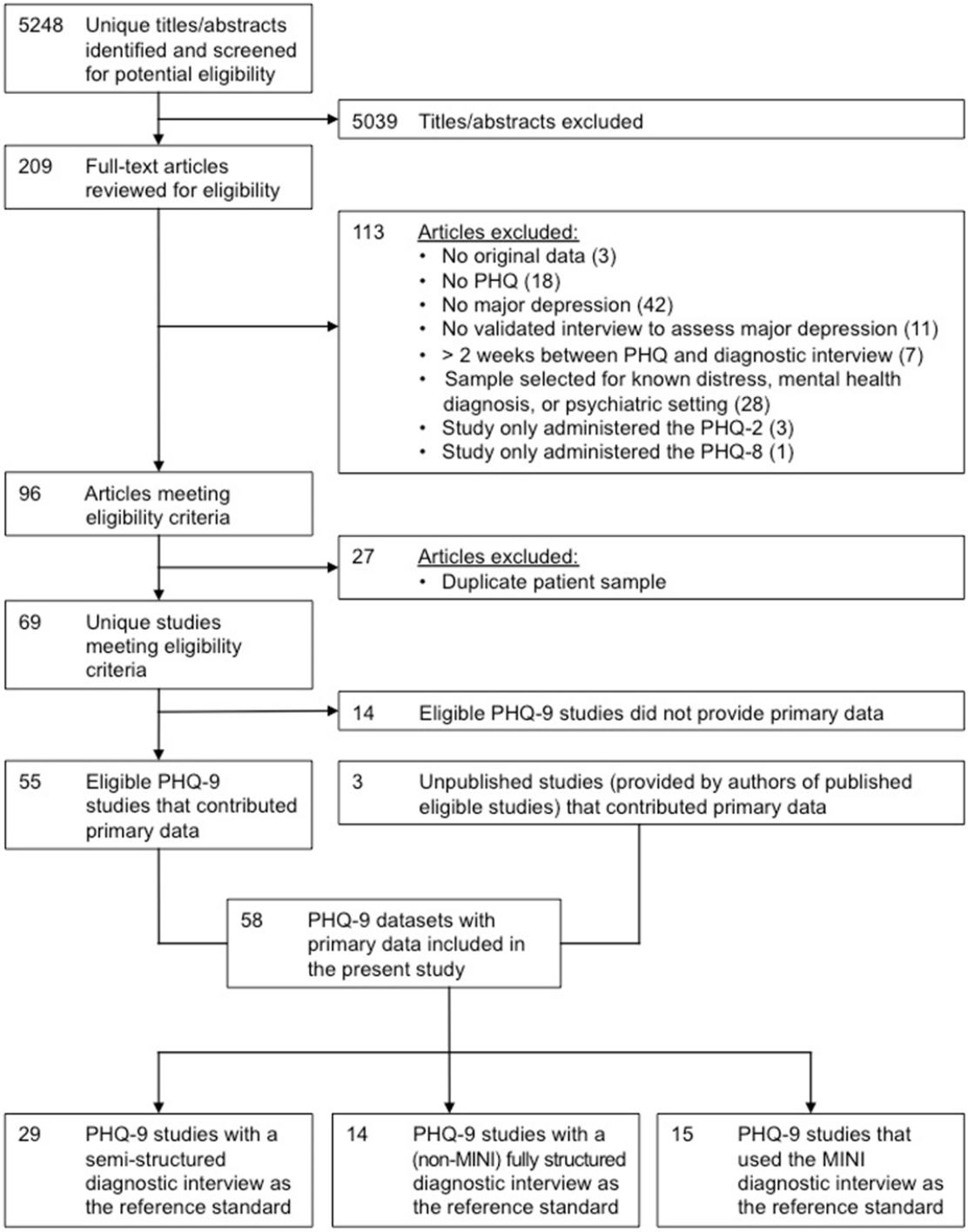
5. **Overall risk of bias:** Rate as "low", "high", or "unclear" risk of bias. Given that questions 2 and 3 will typically be coded as "yes", use the following rules to code the overall risk of bias:

   SQ1 = UNCLEAR and SQ4 = YES: code as UNCLEAR risk of bias
   SQ1 = UNCLEAR and SQ4 = UNCLEAR: code as UNCLEAR risk of bias
   SQ1 = UNCLEAR and SQ4 = NO: code as HIGH risk of bias if the % in SQ4 is <50% and code as UNCLEAR risk of bias if the % in SQ4 is >=50%
   SQ1 = YES and SQ4 = UNCLEAR: code as UNCLEAR risk of bias
   SQ1 = YES and SQ4 = YES: code as LOW risk of bias
   SQ1 = YES and SQ4 = NO: code as HIGH risk of bias if the % in SQ4 is <50% and code as UNCLEAR risk of bias if the % in SQ4 is >=50%

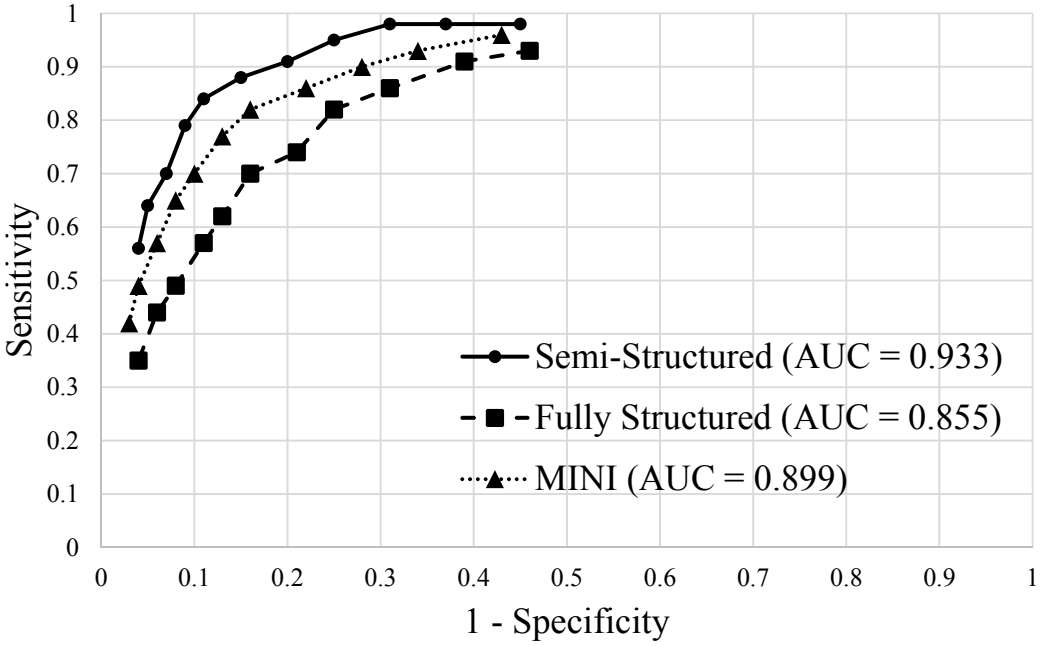
   Note: If "IPD" was selected for signalling question 1, and the overall risk of bias rating depends on the individual patient rating in signalling question 1, then rate as "IPD" and indicate which participants should receive which bias rating (for example, participants administered the reference standard within 1 week are rated as "low", whereas those administered the reference standard within 1-2 weeks are rated as "unclear").

   Please indicate factors in decision in "Notes".

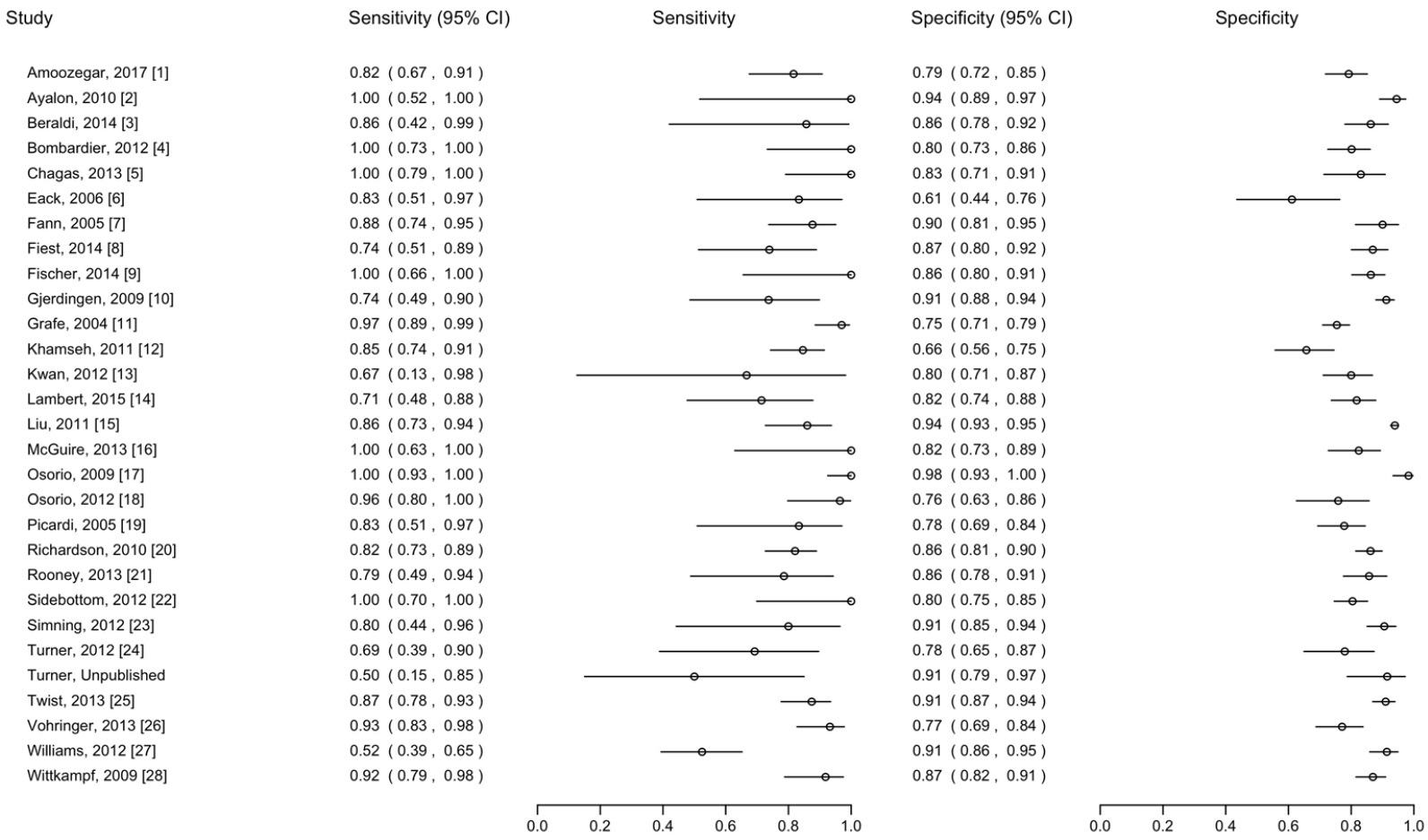
Supplementary figure A. Flow diagram of study selection process



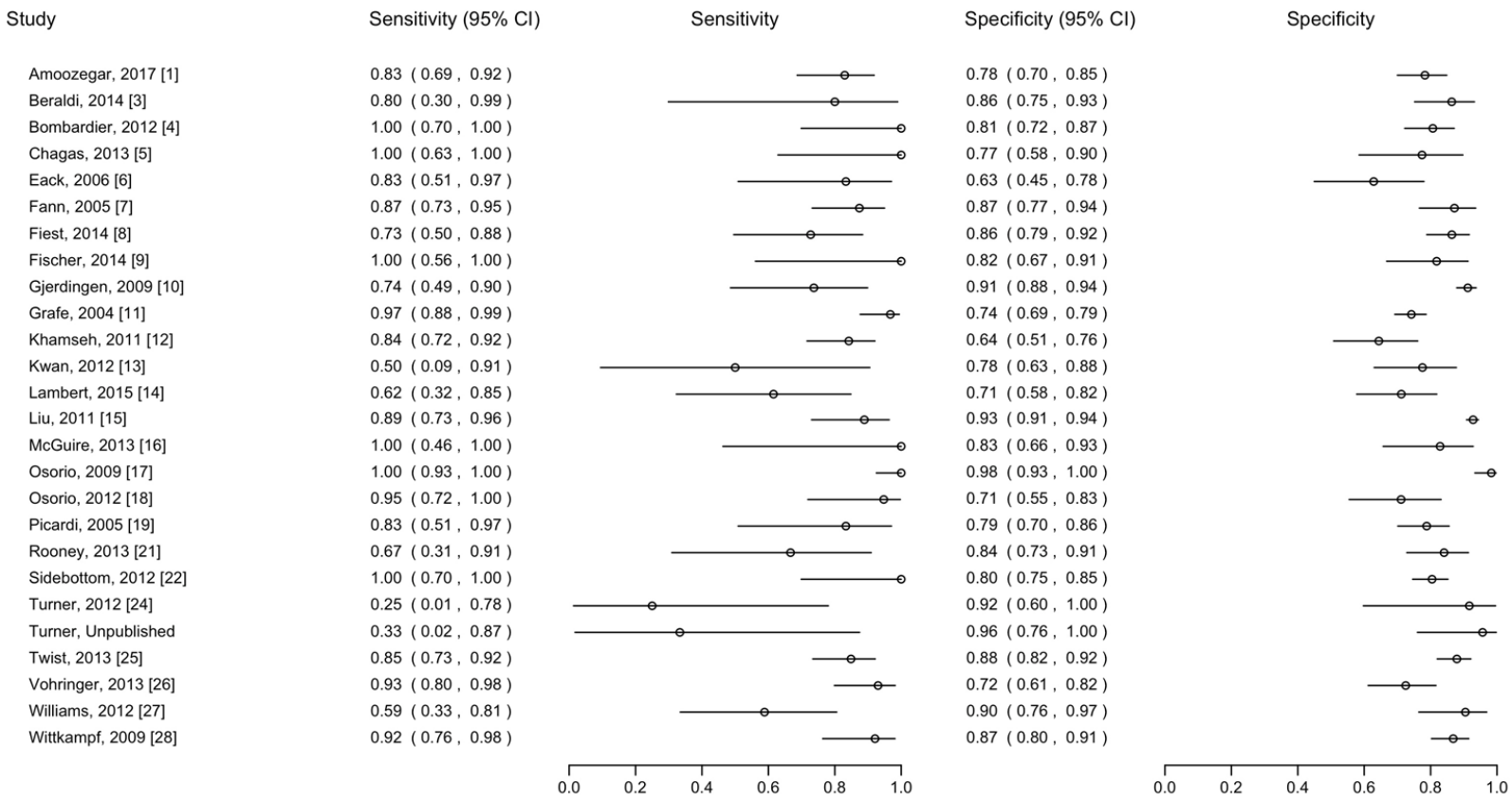
Supplementary figure B. ROC curves comparing sensitivity and specificity estimates for each reference standard category



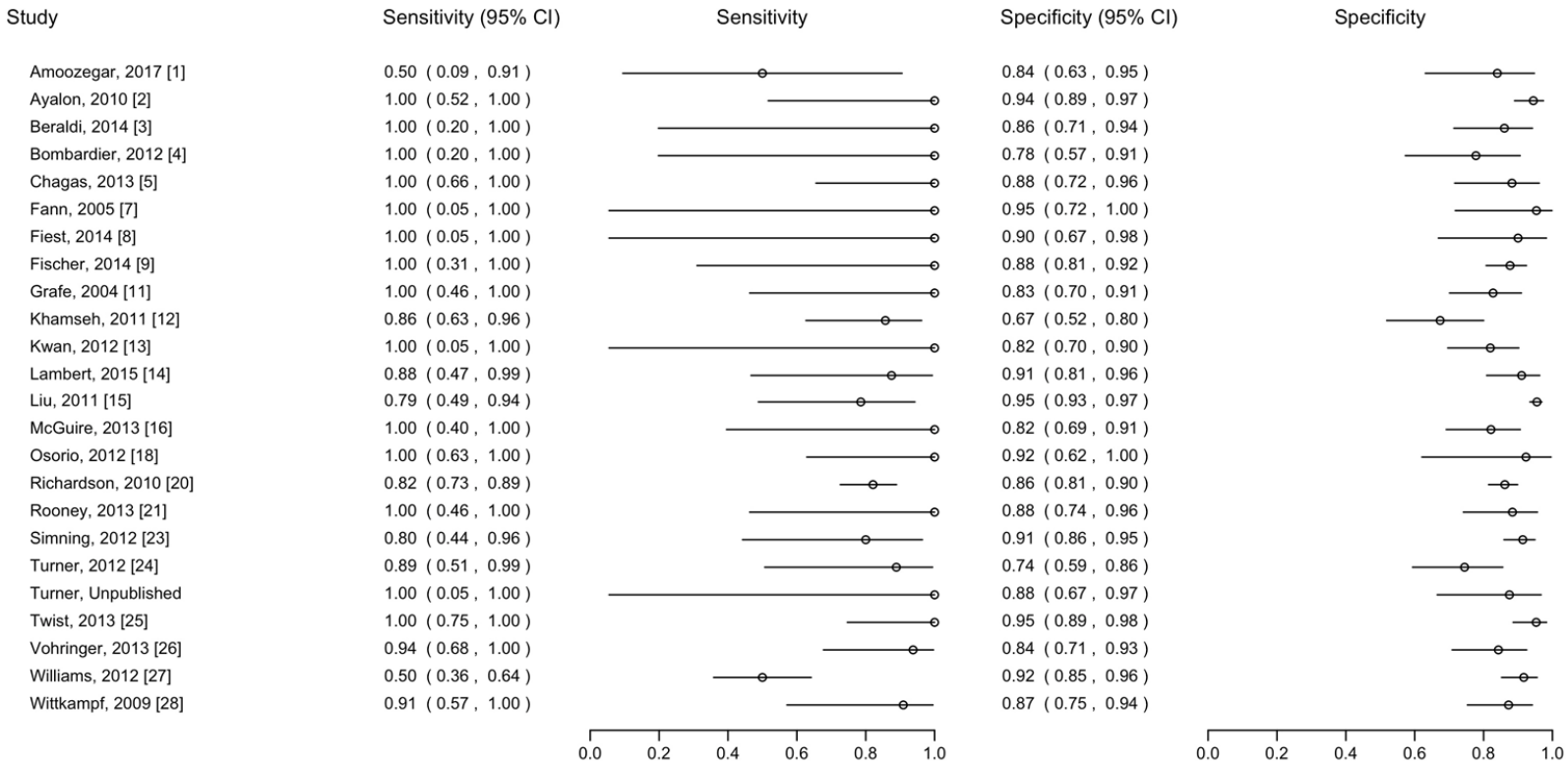
Supplementary figure C1. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 29; N Participants = 6,725; N major depression = 924)



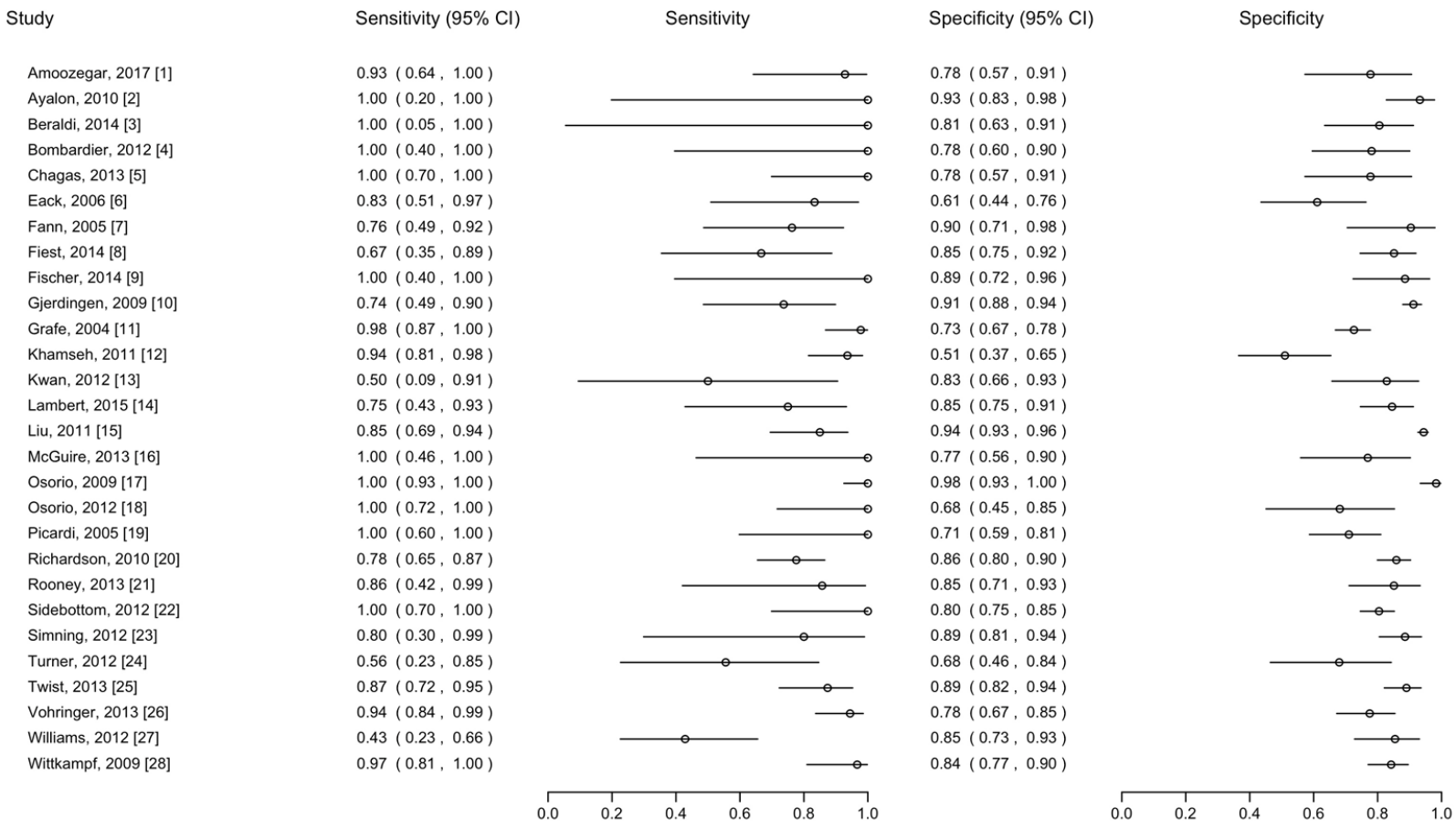
Supplementary figure C2. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants aged <60, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 26; N Participants = 4,132; N major depression = 629)



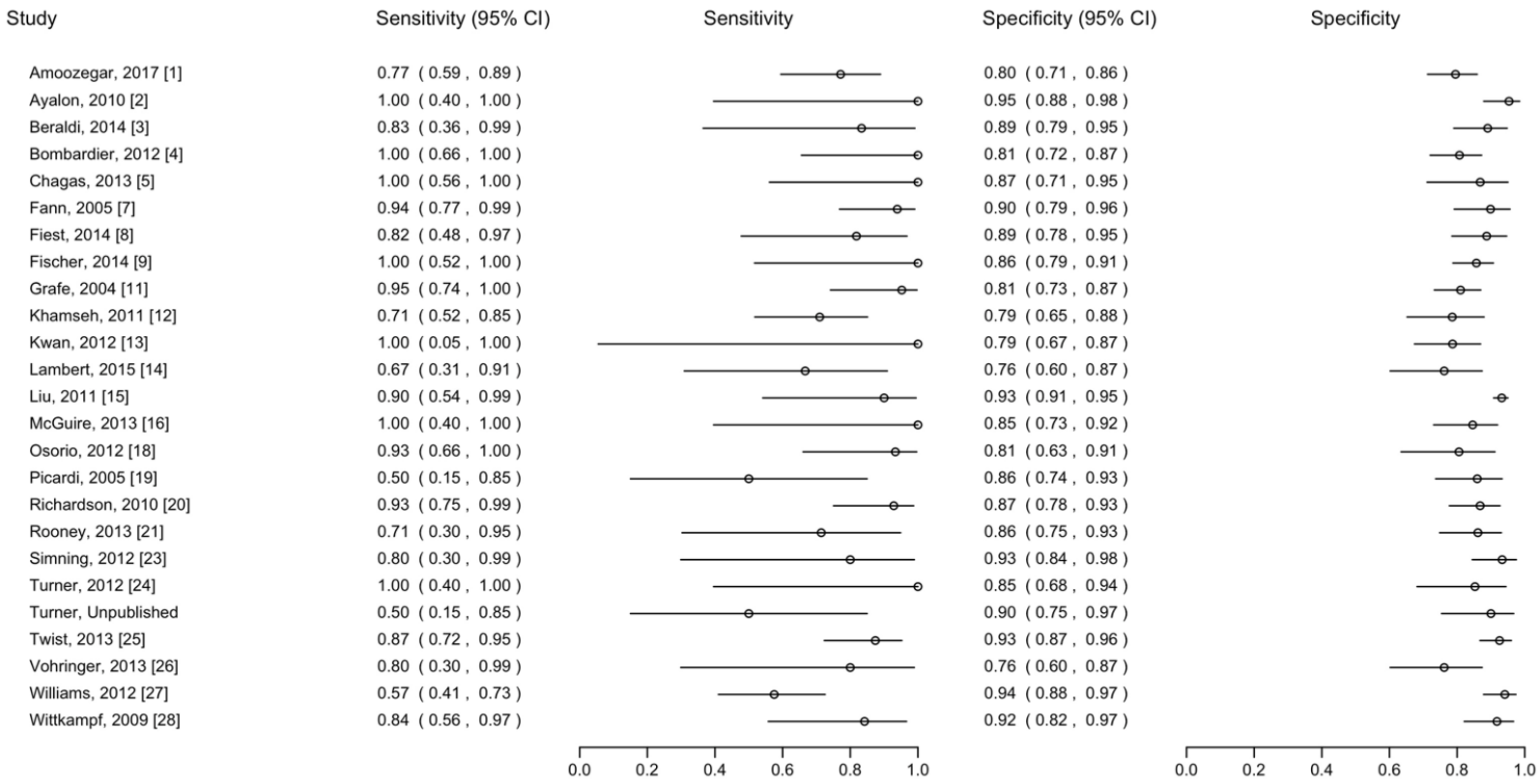
Supplementary figure C3. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants aged ≥60, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 24; N Participants = 2,577; N major depression = 295)



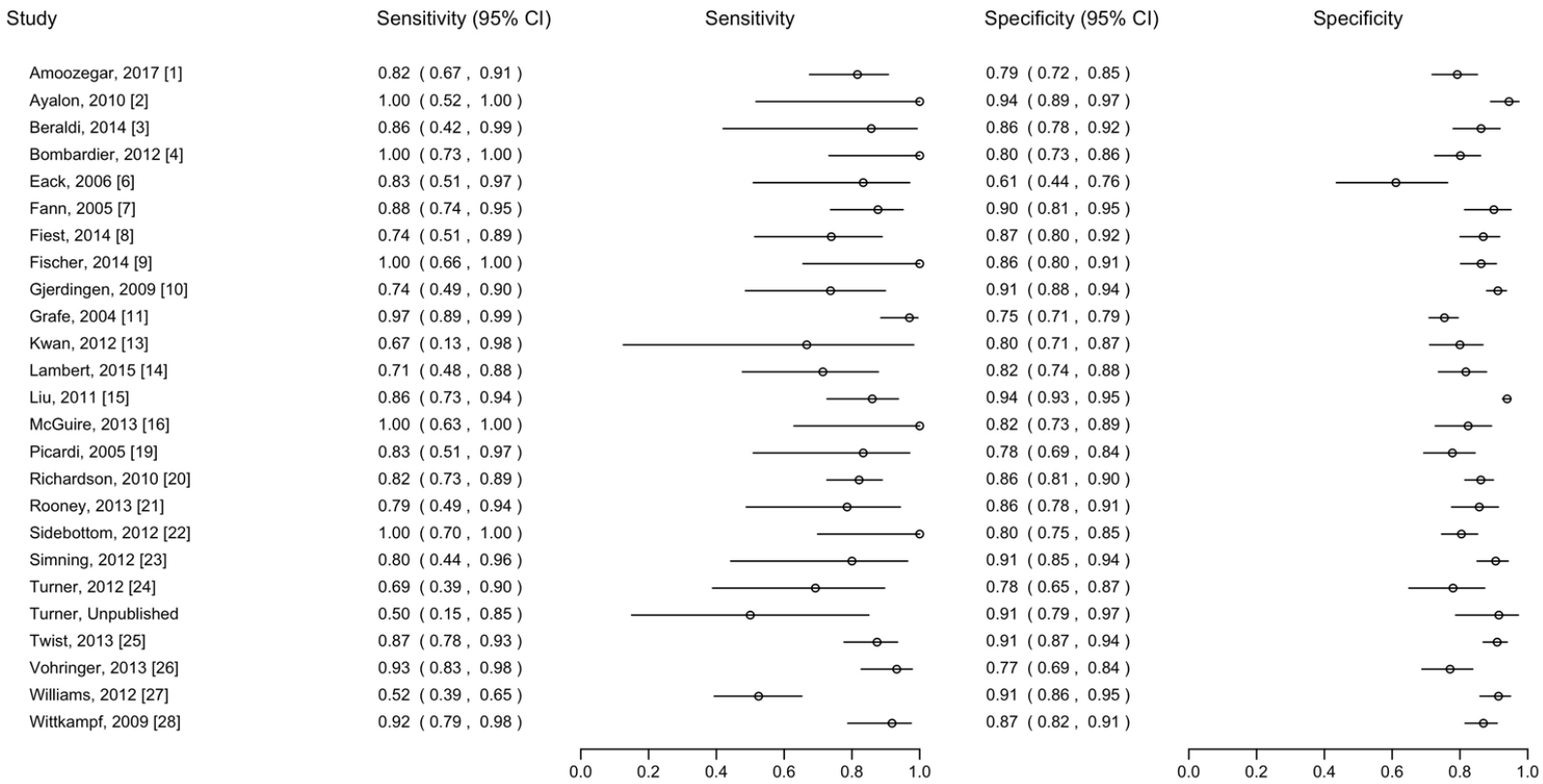
Supplementary figure C4. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among women, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 28; N Participants = 3,906; N major depression = 573)



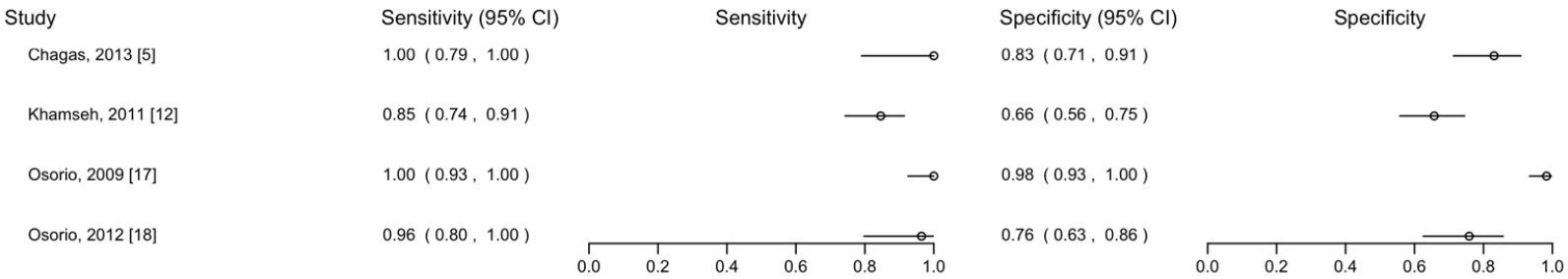
Supplementary figure C5. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among men, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 25; N Participants = 2,812; N major depression = 351)



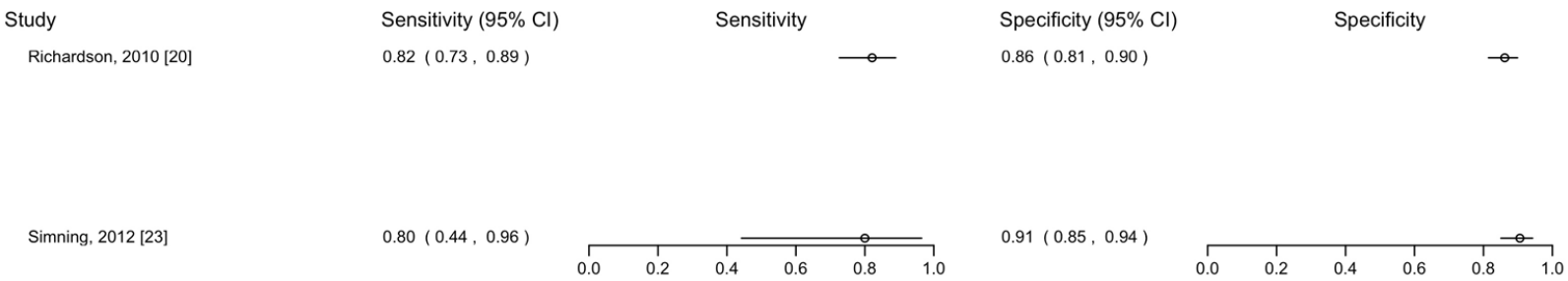
Supplementary figure C6. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a very high human development index, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 25; N Participants = 6,195; N major depression = 739)



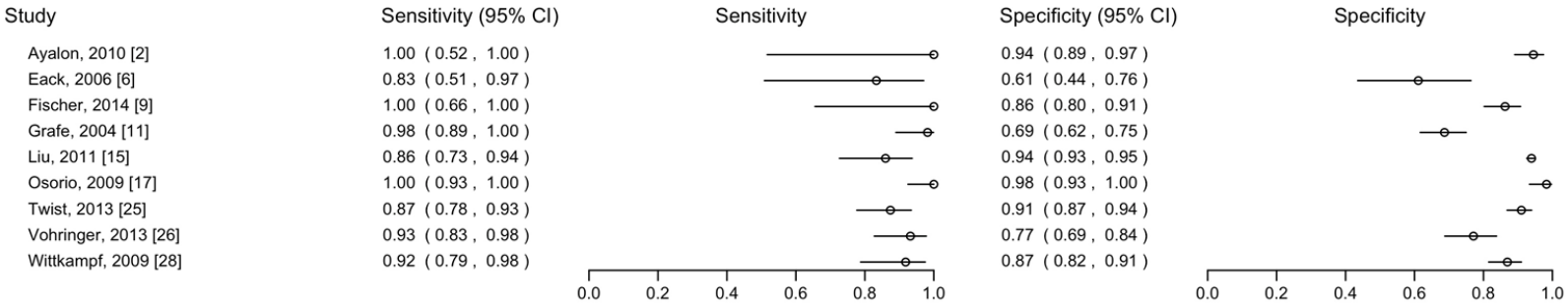
Supplementary figure C7. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a high human development index, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 4; N Participants = 530; N major depression = 185)



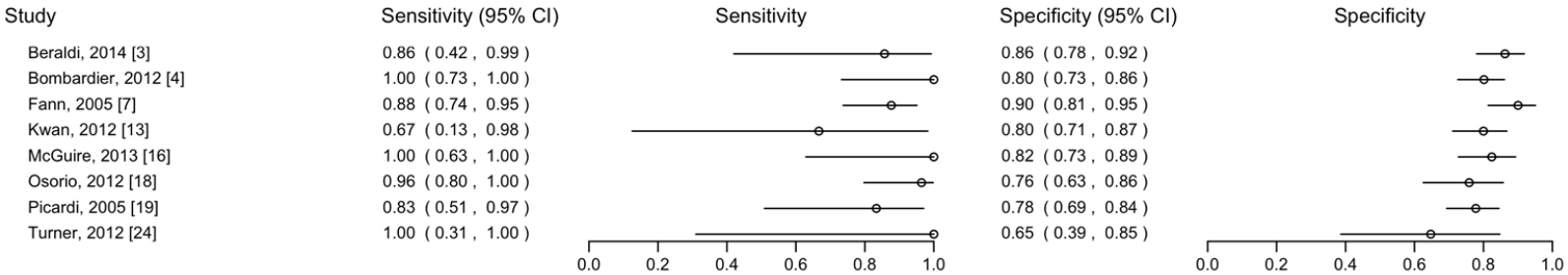
Supplementary figure C8. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a non-medical setting, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 2; N Participants = 567; N major depression = 105)



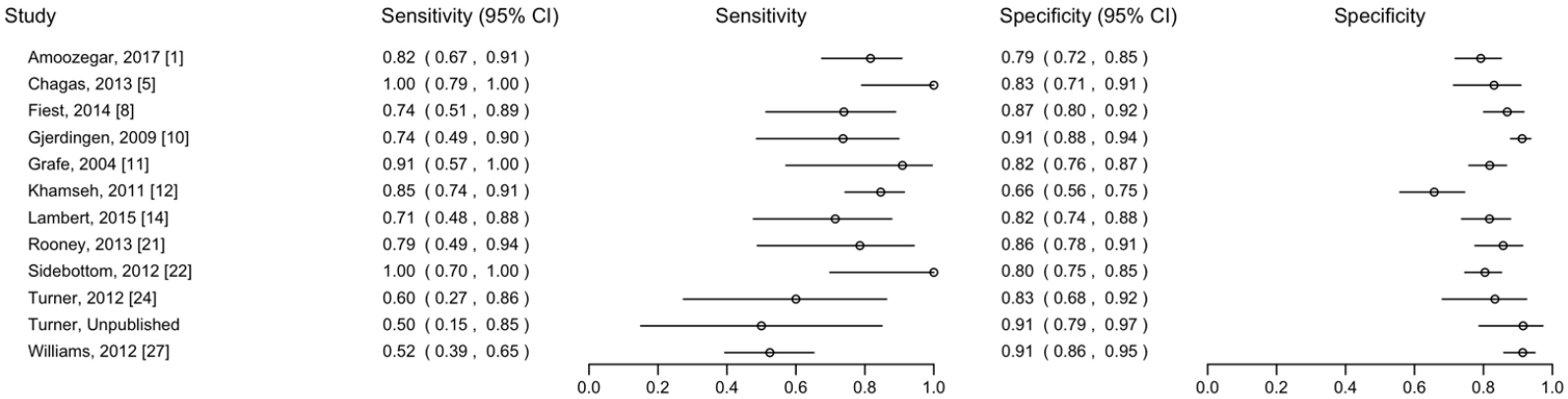
Supplementary figure C9. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a primary care setting, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 9; N Participants = 3,163; N major depression = 377)



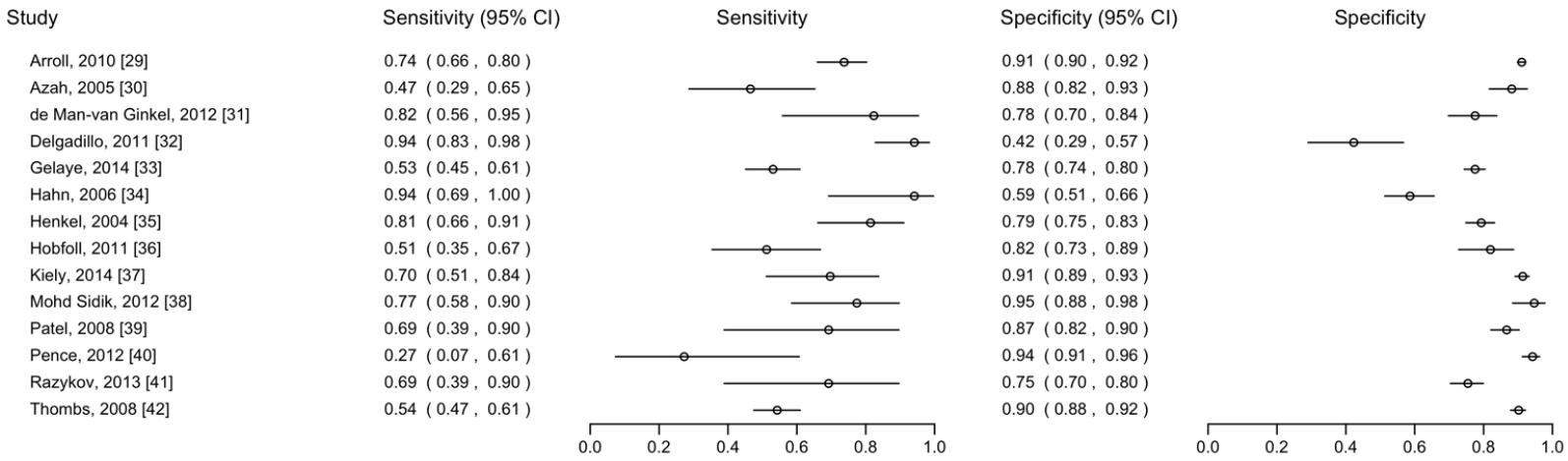
Supplementary figure C10. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from an inpatient specialty care setting, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 8; N Participants = 867; N major depression = 121)



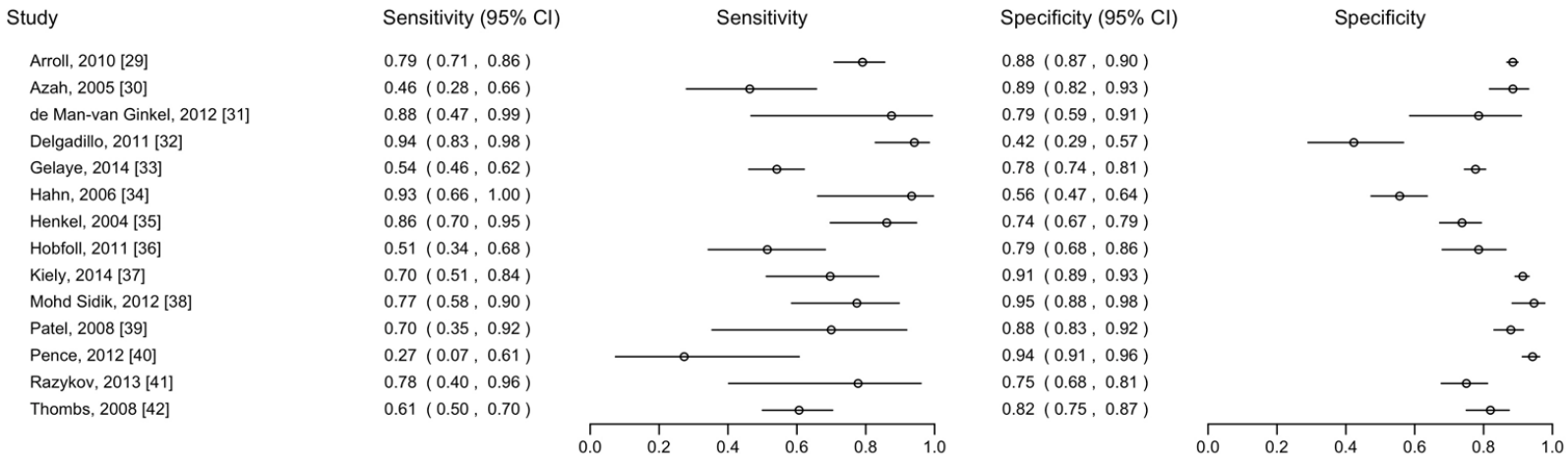
Supplementary figure C11. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from an outpatient specialty care setting, among studies that used a semi-structured diagnostic interview as the reference standard (N Studies = 12; N Participants = 2,128; N major depression = 321)



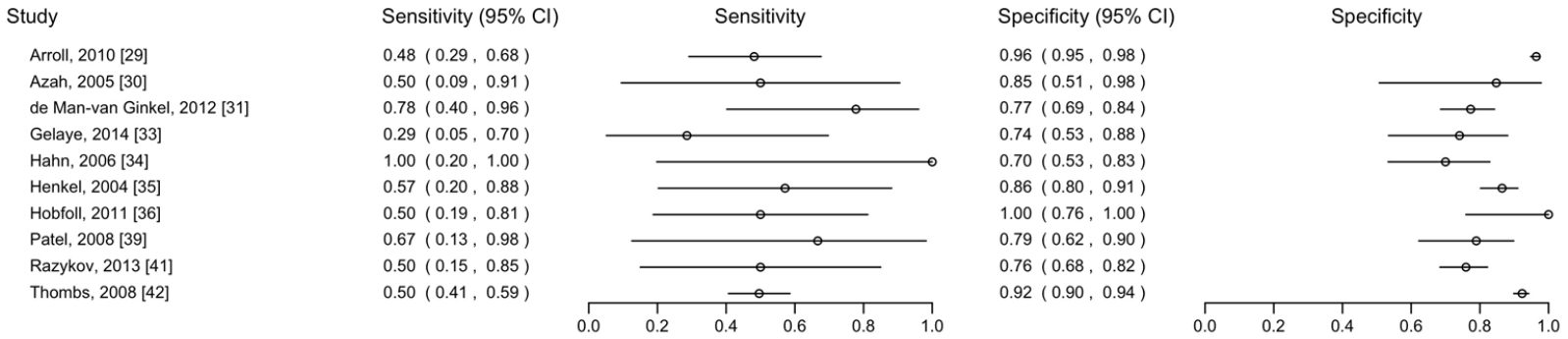
Supplementary figure 12. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 14; N Participants = 7,680; N major depression = 839)



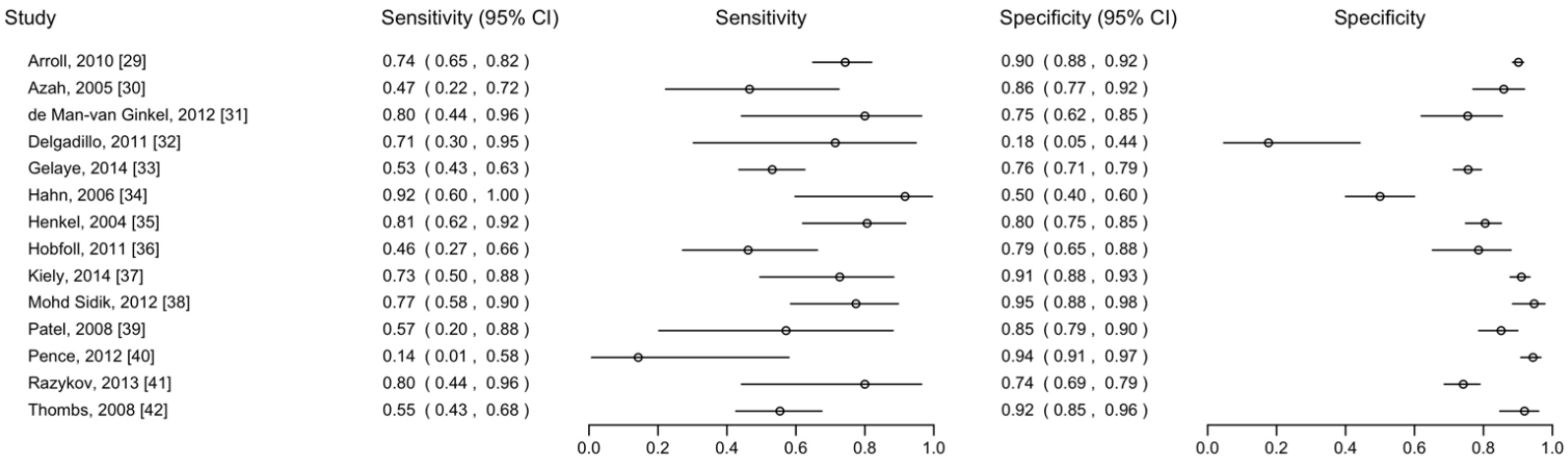
Supplementary figure 13. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants aged <60, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 14; N Participants = 5,504; N major depression = 645)



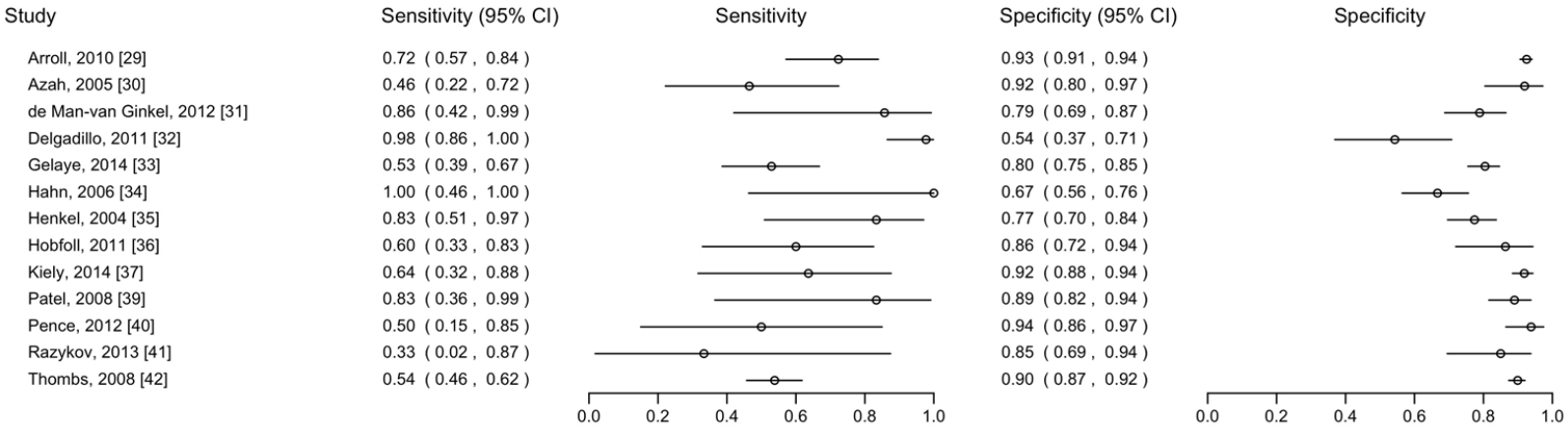
Supplementary figure C14. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants aged ≥60, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 10; N Participants = 2,175; N major depression = 194)



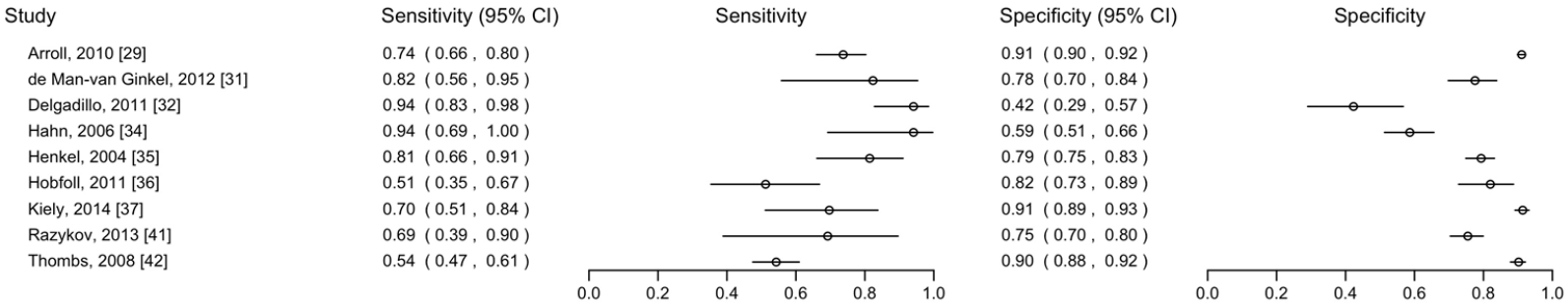
Supplementary figure C15. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among women, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 14; N Participants = 4,285; N major depression = 463)



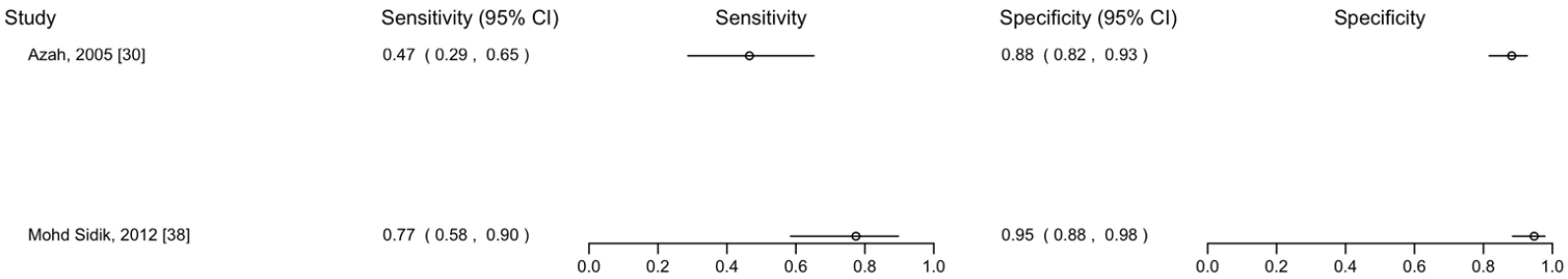
Supplementary figure C16. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among men, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 13; N Participants = 3,395; N major depression = 376)



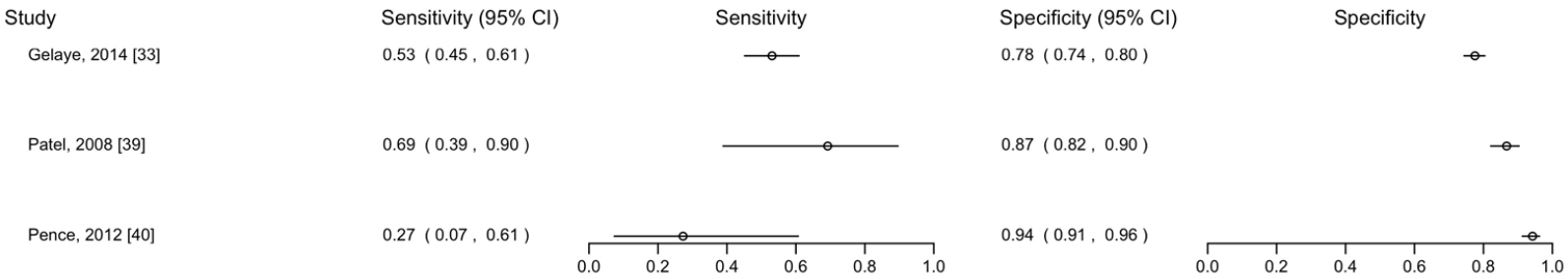
Supplementary figure C17. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a very high human development index, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 9; N Participants = 5,740; N major depression = 592)



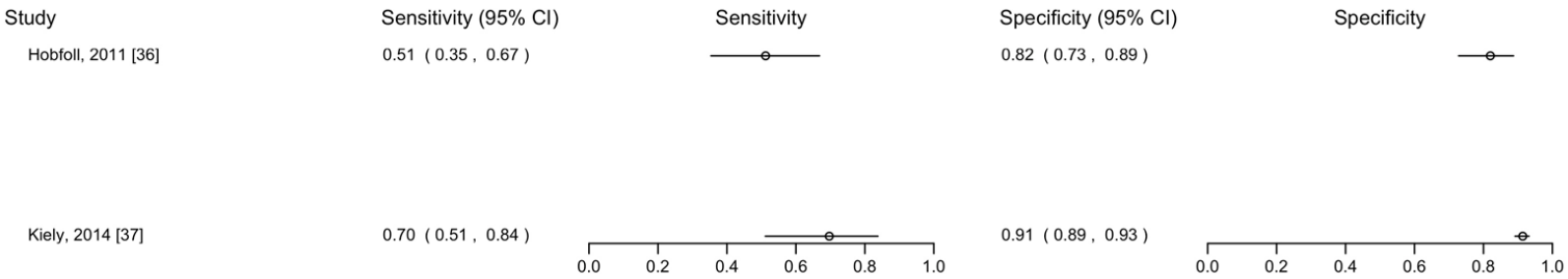
Supplementary figure C18. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a high human development index, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 2; N Participants = 326; N major depression = 61)



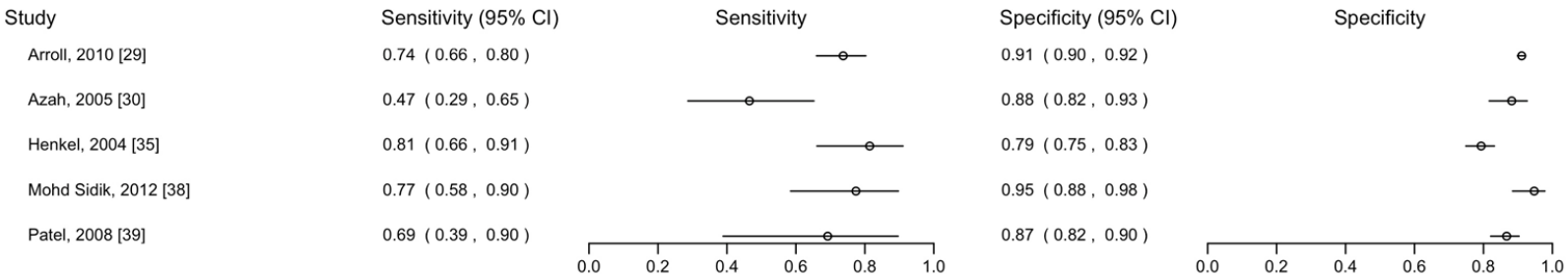
Supplementary figure C19. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a low-medium human development index, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 3; N Participants = 1,614; N major depression = 186)



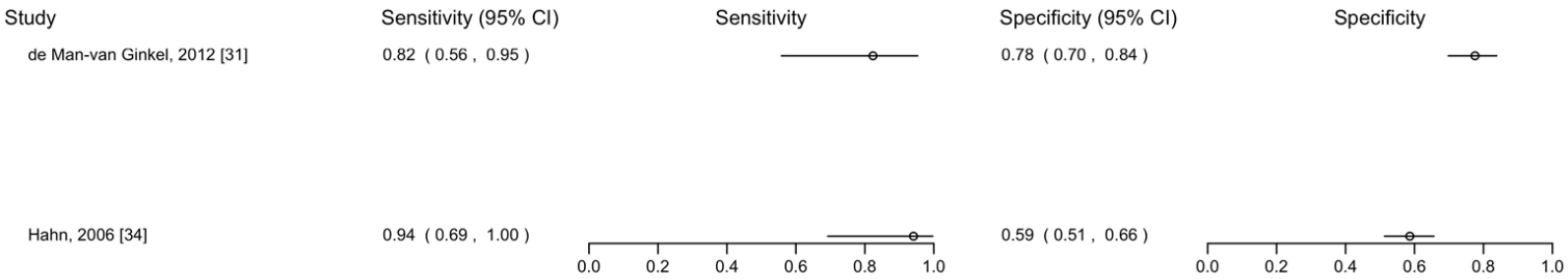
Supplementary figure C20. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a non-medical setting, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 2; N Participants = 963; N major depression = 74)



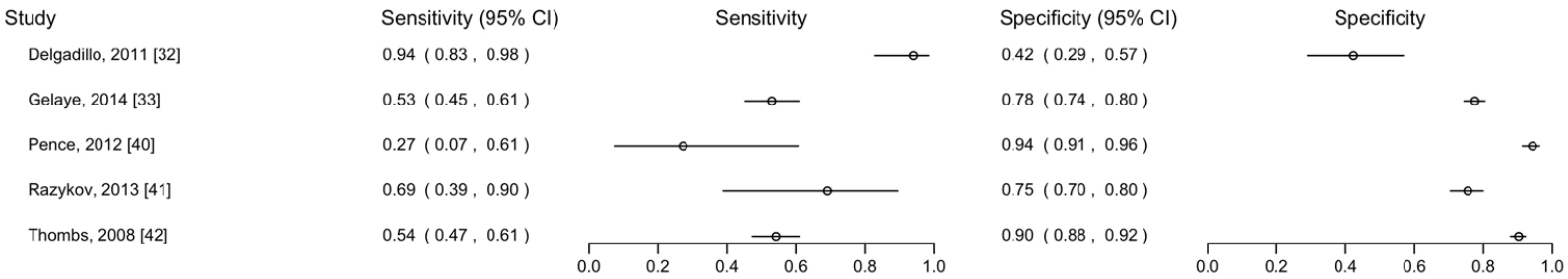
Supplementary figure C21. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a primary care setting, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 5; N Participants = 3,578; N major depression = 273)



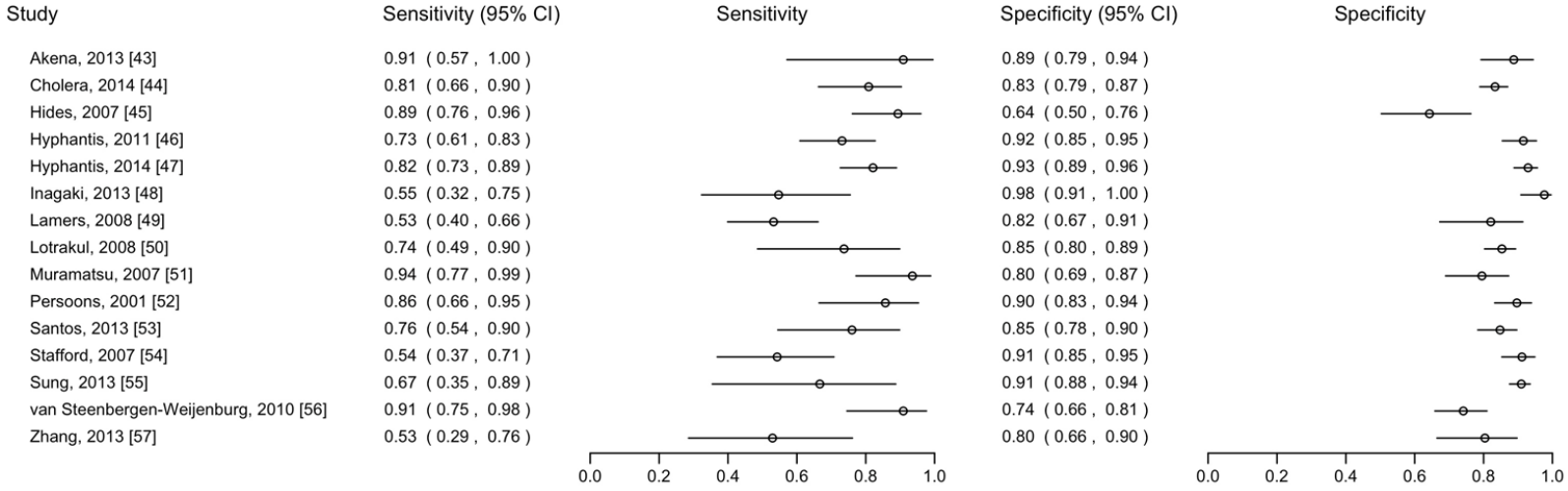
Supplementary figure C22. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from an inpatient specialty care setting, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 2; N Participants = 372; N major depression = 34)



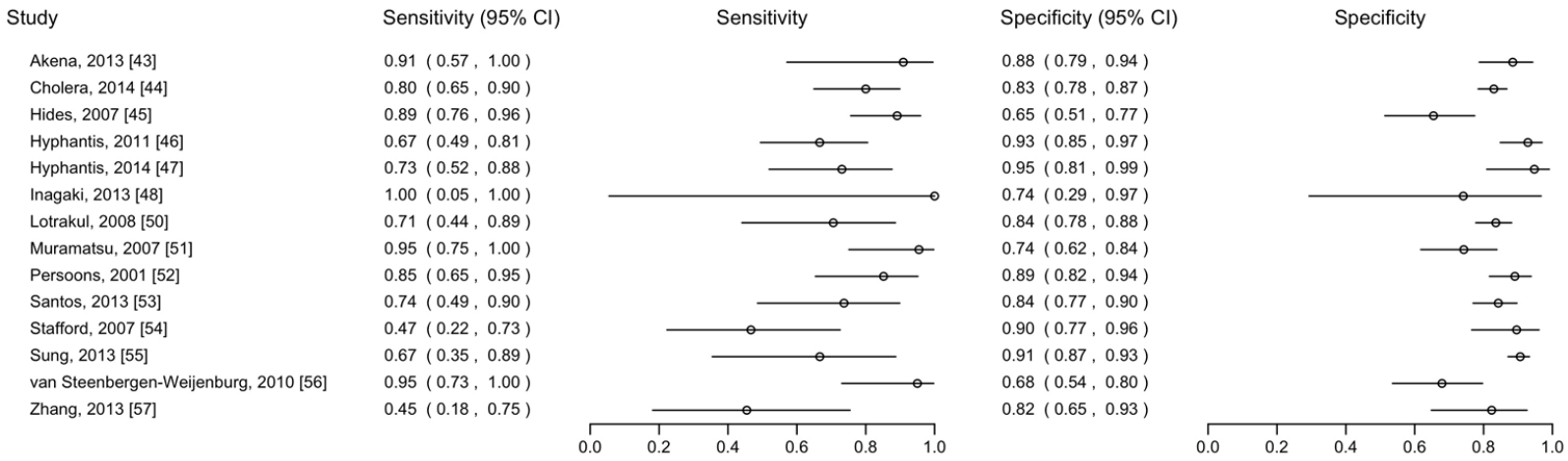
Supplementary figure C23. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from an outpatient specialty care setting, among studies that used a fully structured diagnostic interview as the reference standard (N Studies = 5; N Participants = 2,767; N major depression = 458)



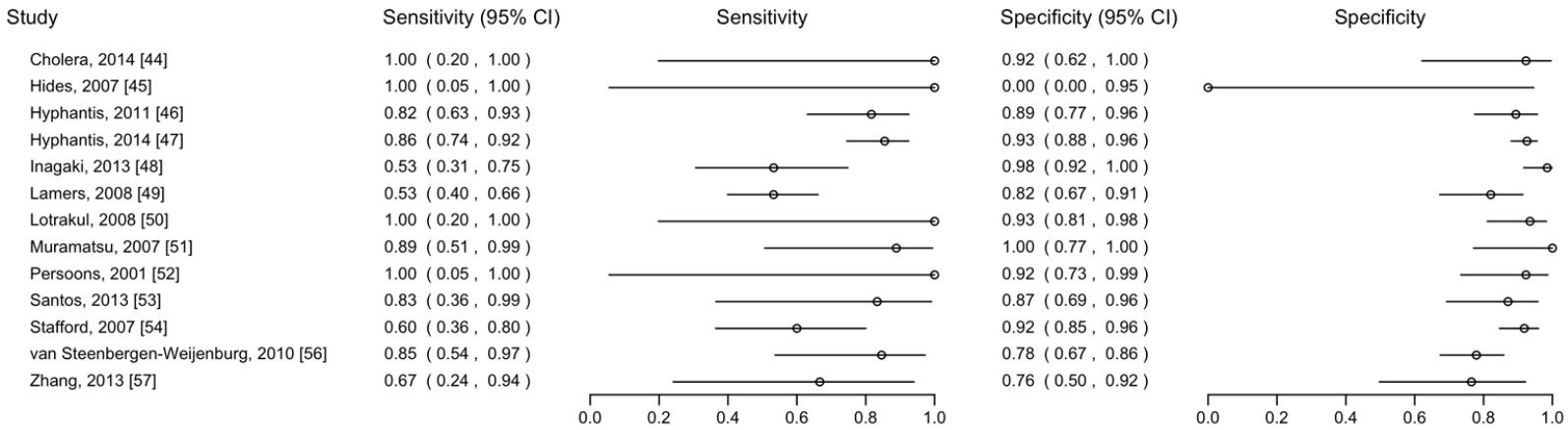
Supplementary figure C24. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9, among studies that used the MINI as the reference standard (N Studies = 15; N Participants = 2,952; N major depression = 549)



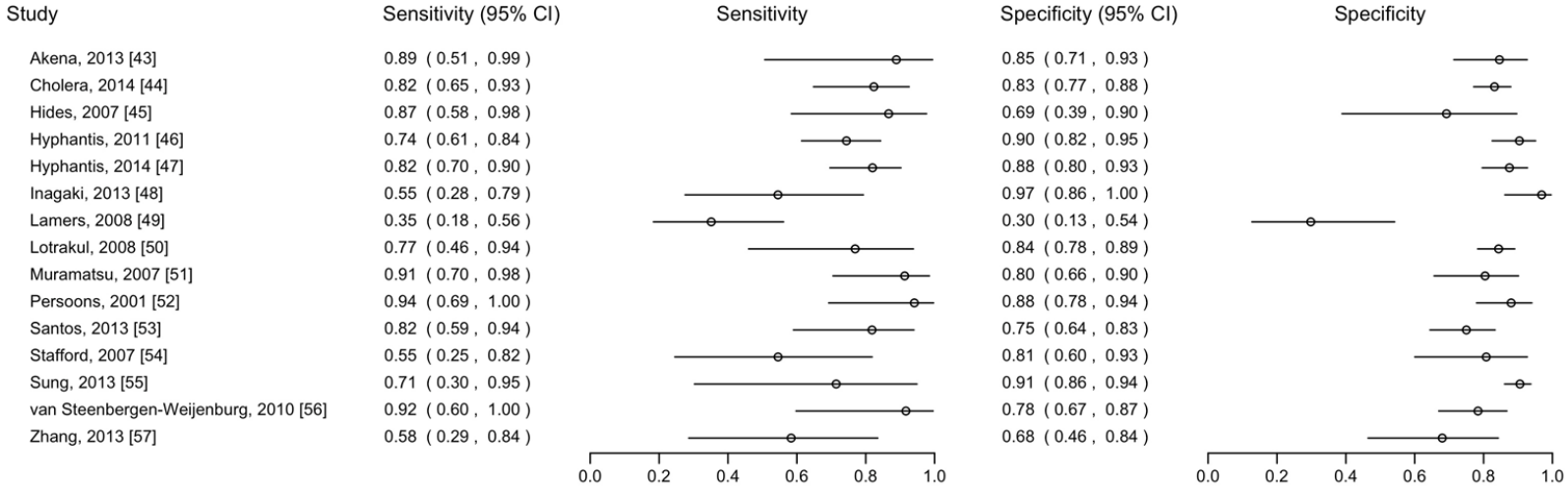
Supplementary figure C25. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants aged <60, among studies that used the MINI as the reference standard (N Studies = 14; N Participants = 1,958; N major depression = 310)



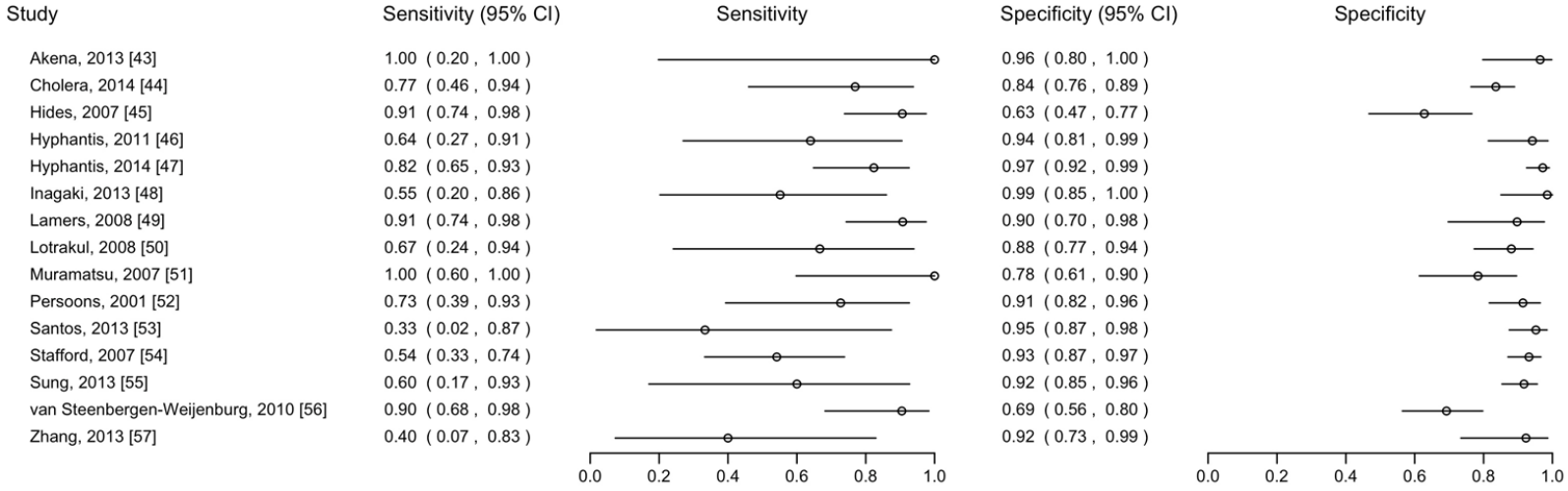
Supplementary figure C26. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants aged ≥60, among studies that used the MINI as the reference standard (N Studies = 13; N Participants = 979; N major depression = 239)



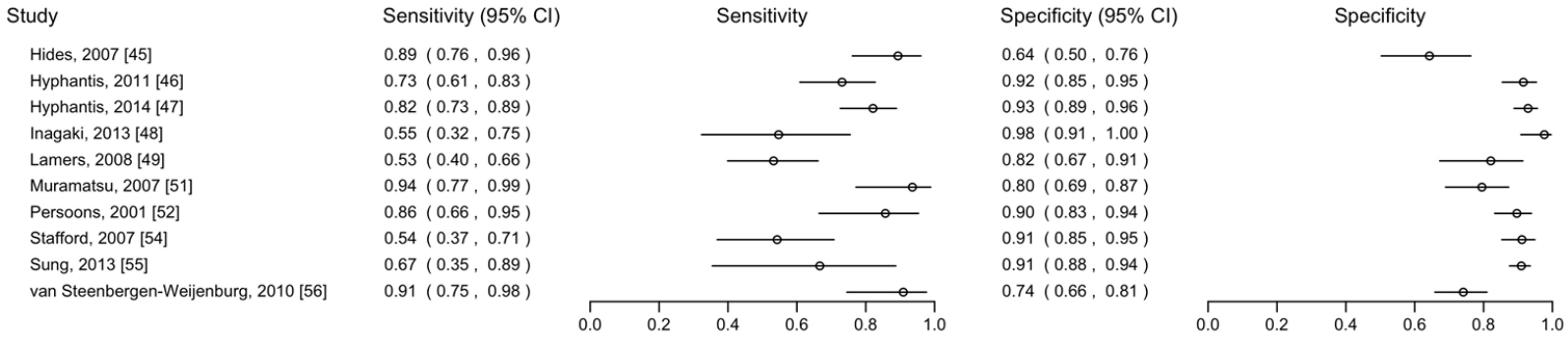
Supplementary figure C27. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among women, among studies that used the MINI as the reference standard (N Studies = 15; N Participants = 1,666; N major depression = 337)



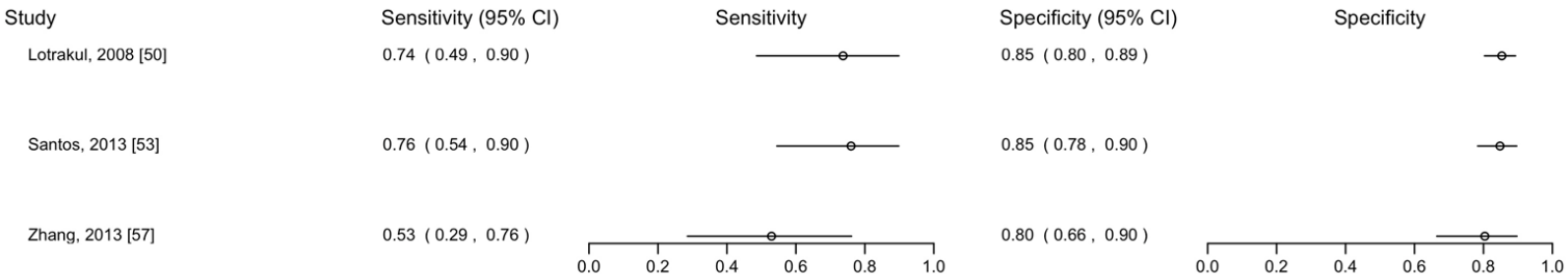
Supplementary figure C28. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among men, among studies that used the MINI as the reference standard (N Studies = 15; N Participants = 1,286; N major depression = 212)



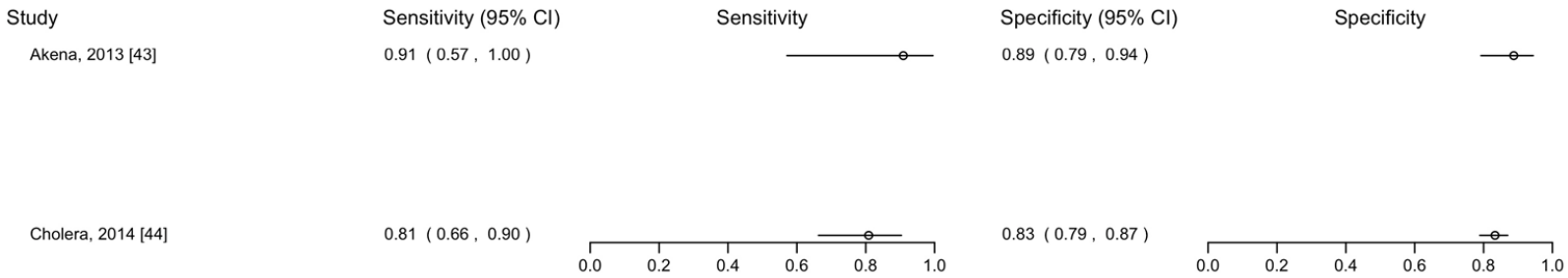
Supplementary figure C29. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a very high human development index, among studies that used the MINI as the reference standard (N Studies = 10; N Participants = 1,924; N major depression = 430)



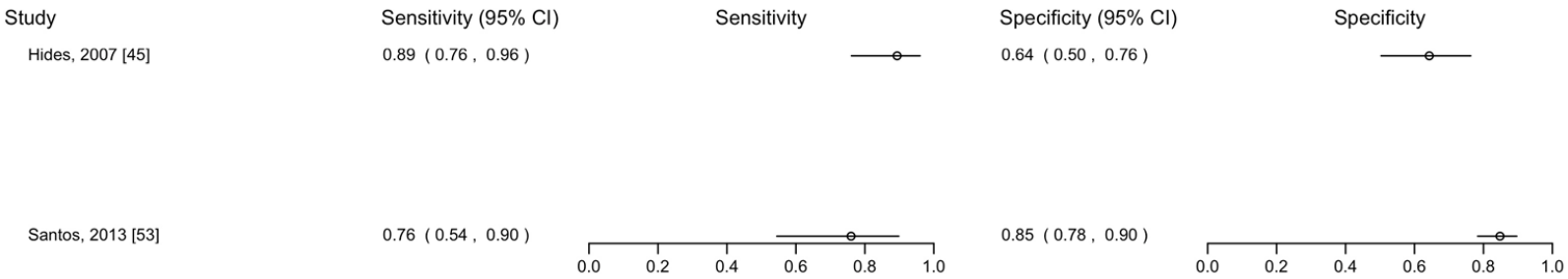
Supplementary figure C30. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a high human development index, among studies that used the MINI as the reference standard (N Studies = 3; N Participants = 542; N major depression = 61)



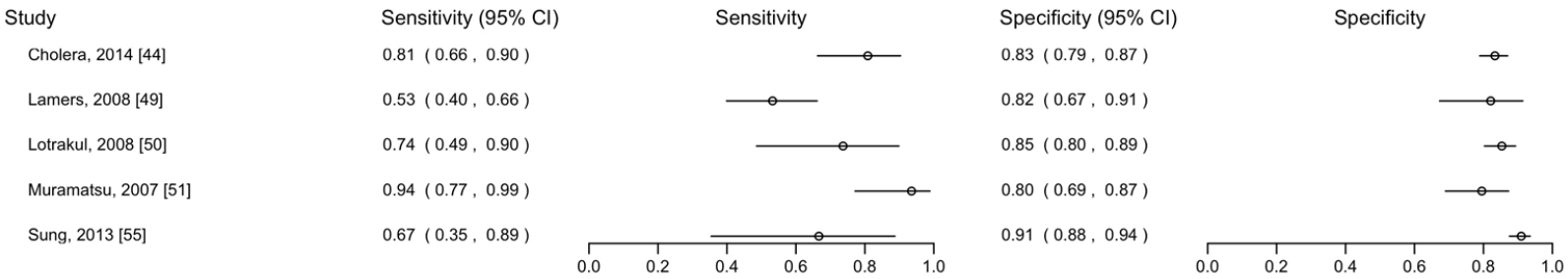
Supplementary figure C31. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a country with a low-medium human development index, among studies that used the MINI as the reference standard (N Studies = 2; N Participants = 486; N major depression = 58)



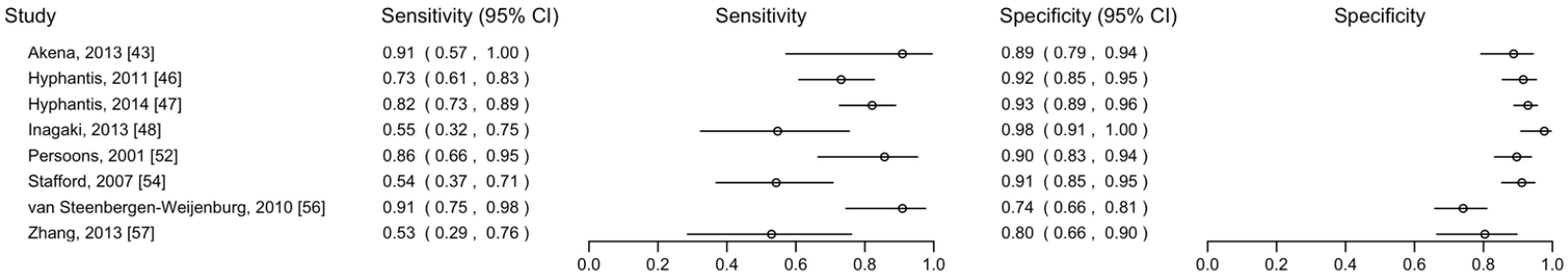
Supplementary figure C32. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a non-medical setting, among studies that used the MINI as the reference standard (N Studies = 2; N Participants = 299; N major depression = 72)



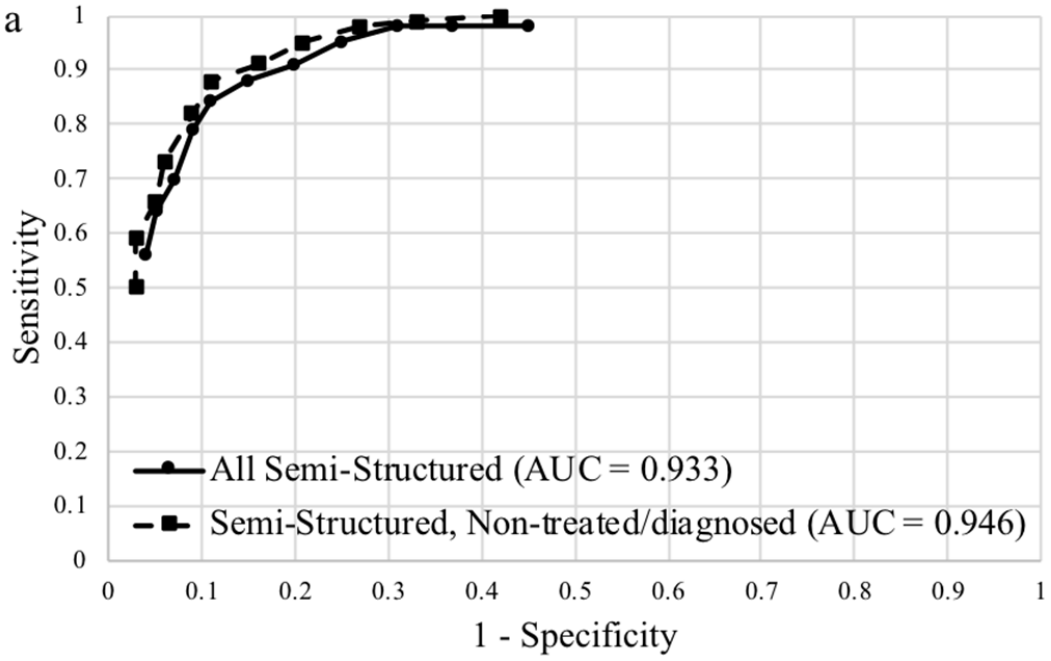
Supplementary figure C32. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a primary care setting, among studies that used the MINI as the reference standard (N Studies = 5; N Participants = 1,290; N major depression = 168)



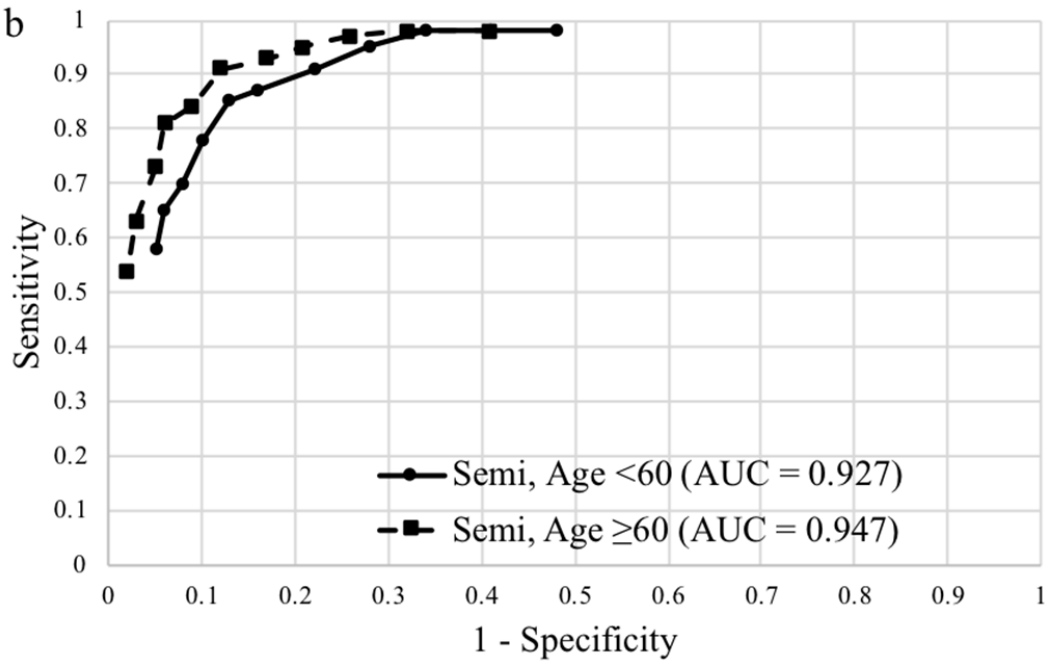
Supplementary figure C33. Forest plots of sensitivity and specificity estimates for cutoff 10 of the PHQ-9 among participants from a specialty care setting, among studies that used the MINI as the reference standard (N Studies = 8; N Participants = 1363; N major depression = 309)



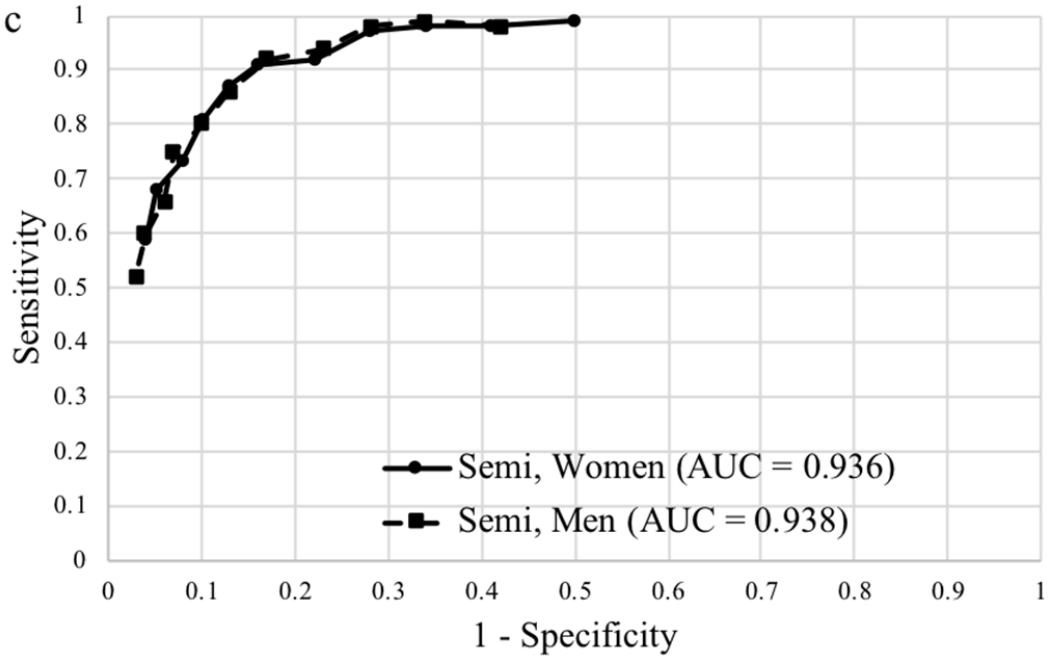
Supplementary figure D1. ROC curves comparing PHQ-9 sensitivity and specificity among all participants compared to participants not currently diagnosed or receiving treatment for a mental health problem, among studies that used a semi-structured diagnostic interview as the reference standard



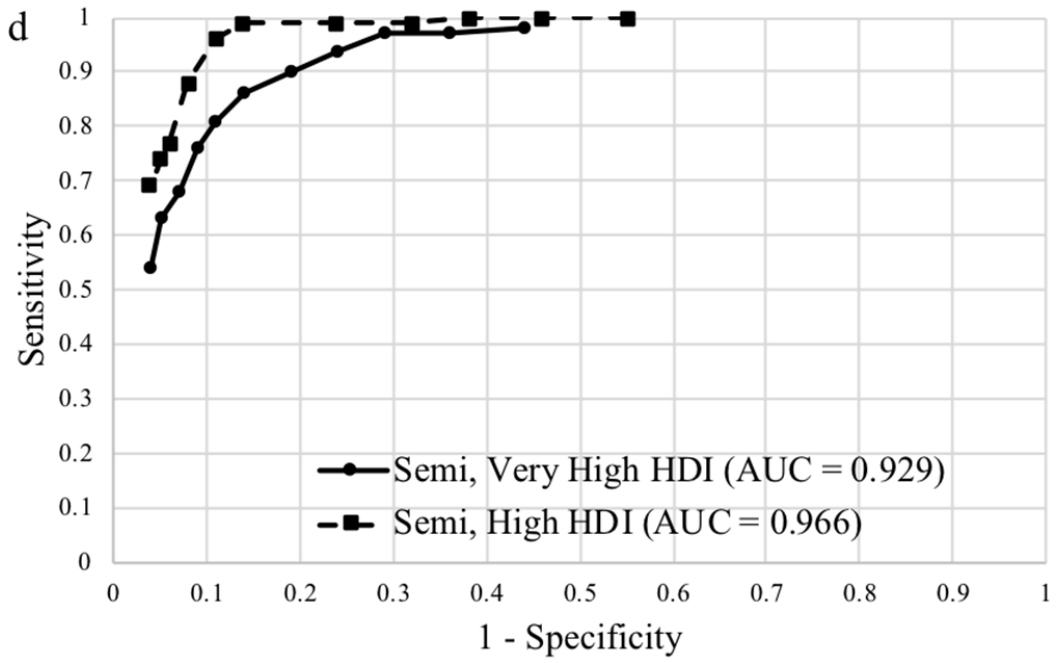
Supplementary figure D2. ROC curves comparing PHQ-9 sensitivity and specificity among among participants aged <60 compared to participants aged ≥60, among studies that used a semi-structured diagnostic interview as the reference standard



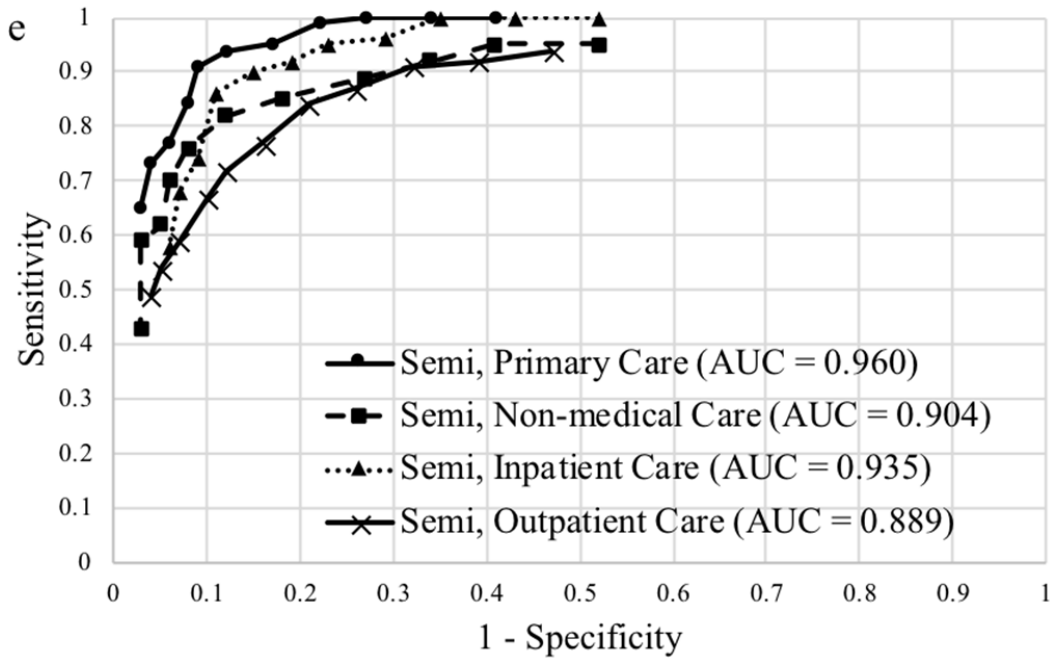
Supplementary figure D3. ROC curves comparing PHQ-9 sensitivity and specificity among among women compared to men, among studies that used a semi-structured diagnostic interview as the reference standard



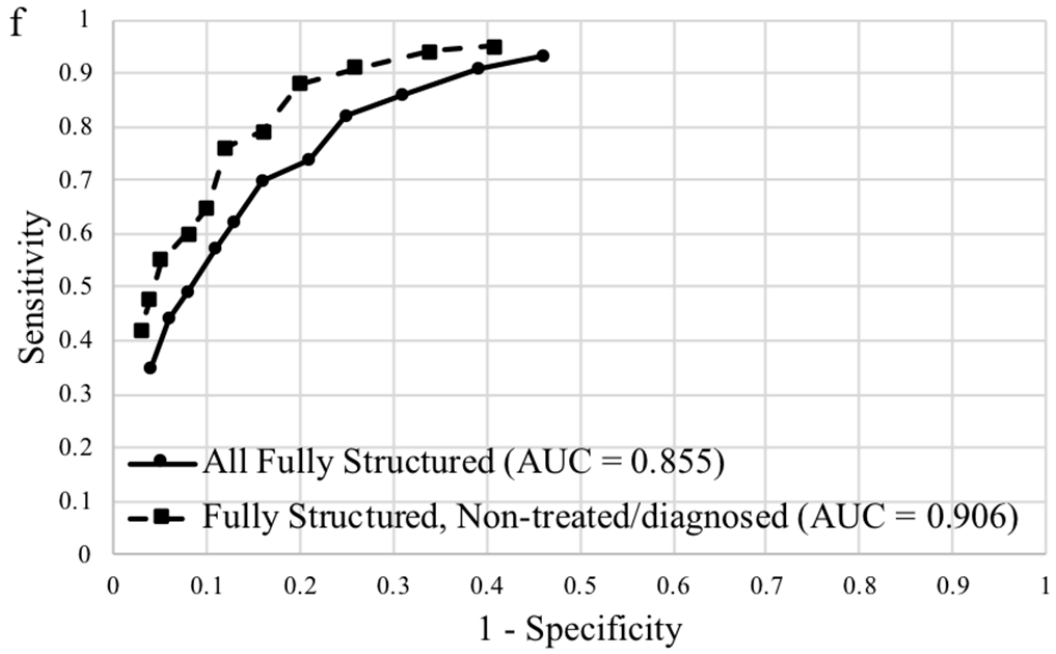
Supplementary figure D4. ROC curves comparing PHQ-9 sensitivity and specificity among participants from countries with a very high human development index compared to a high human development index, among studies that used a semi-structured diagnostic interview as the reference standard



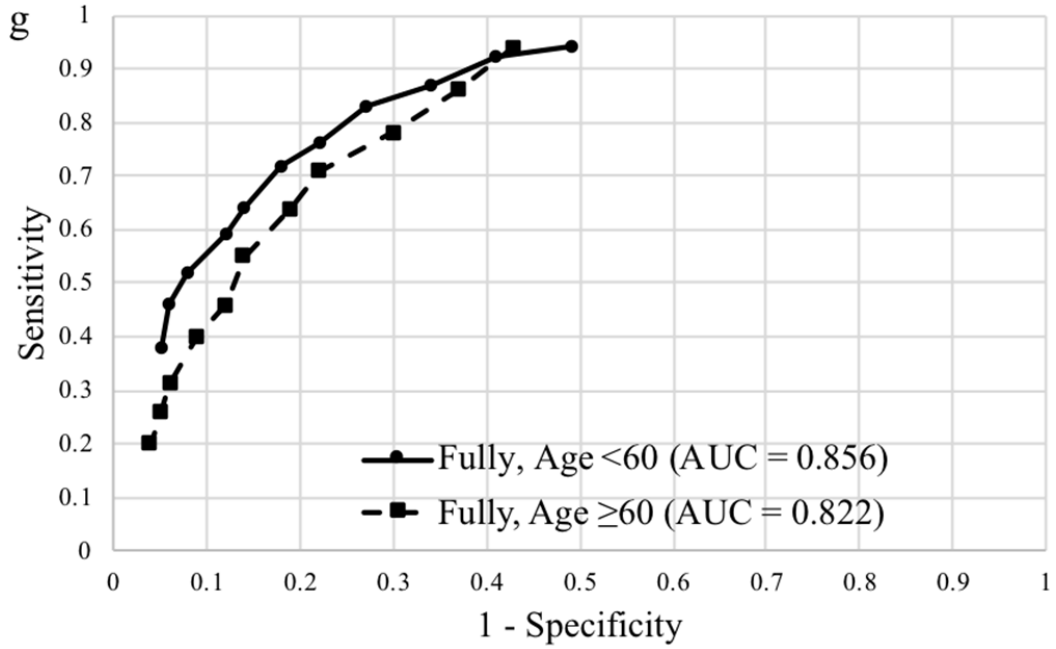
Supplementary figure D5. ROC curves comparing PHQ-9 sensitivity and specificity among participants from non-medical, primary care, inpatient speciality care and outpatient specialty care, among studies that used a semi-structured diagnostic interview as the reference standard



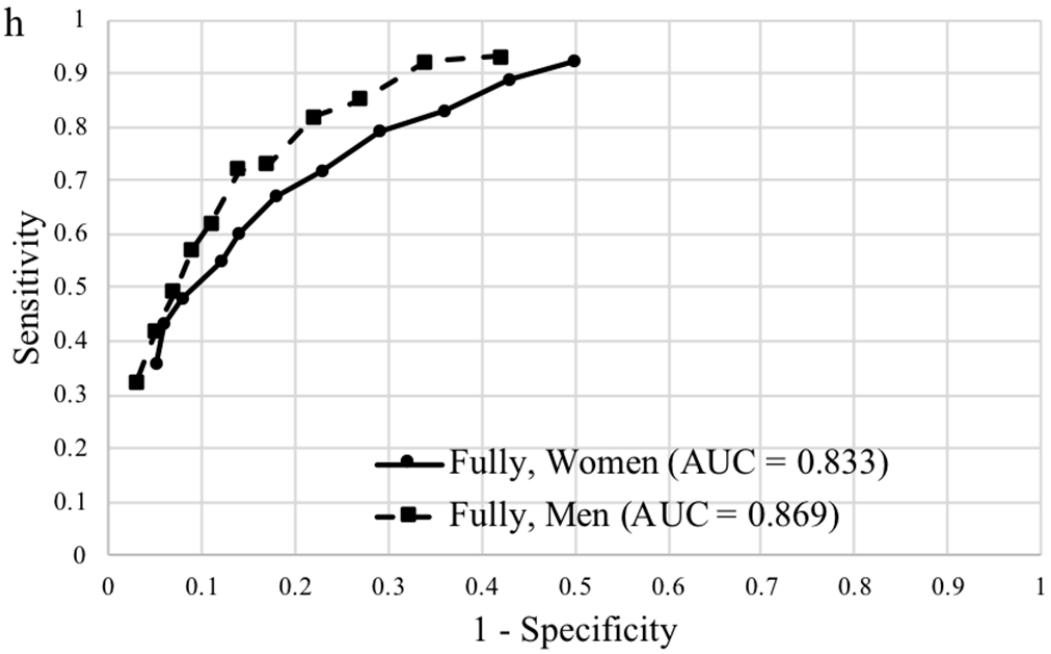
Supplementary figure D6. ROC curves comparing PHQ-9 sensitivity and specificity among all participants compared to participants not currently diagnosed or receiving treatment for a mental health problem, among studies that used a fully structured diagnostic interview as the reference standard



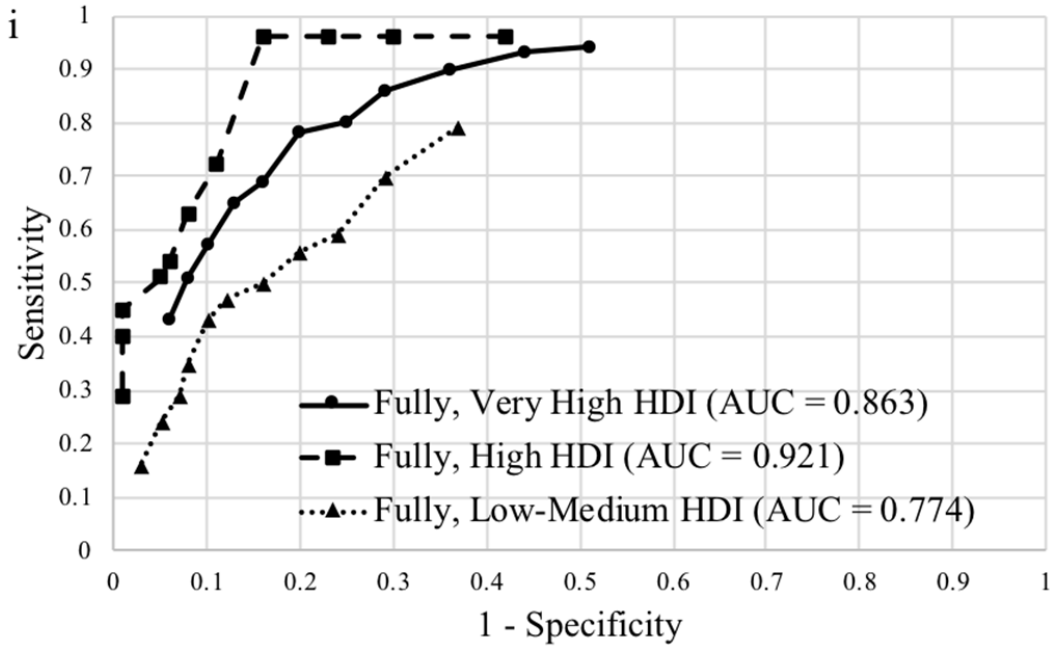
Supplementary figure D7. ROC curves comparing PHQ-9 sensitivity and specificity among among participants aged <60 compared to participants aged ≥60, among studies that used a fully structured diagnostic interview as the reference standard



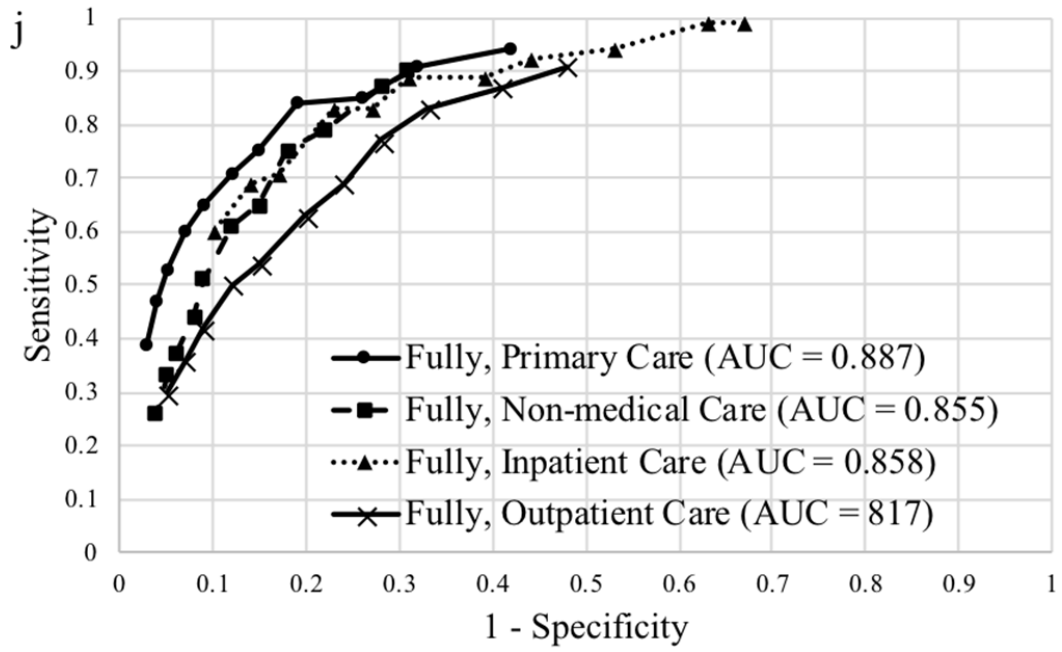
Supplementary figure D8. ROC curves comparing PHQ-9 sensitivity and specificity among among women compared to men, among studies that used a fully structured diagnostic interview as the reference standard



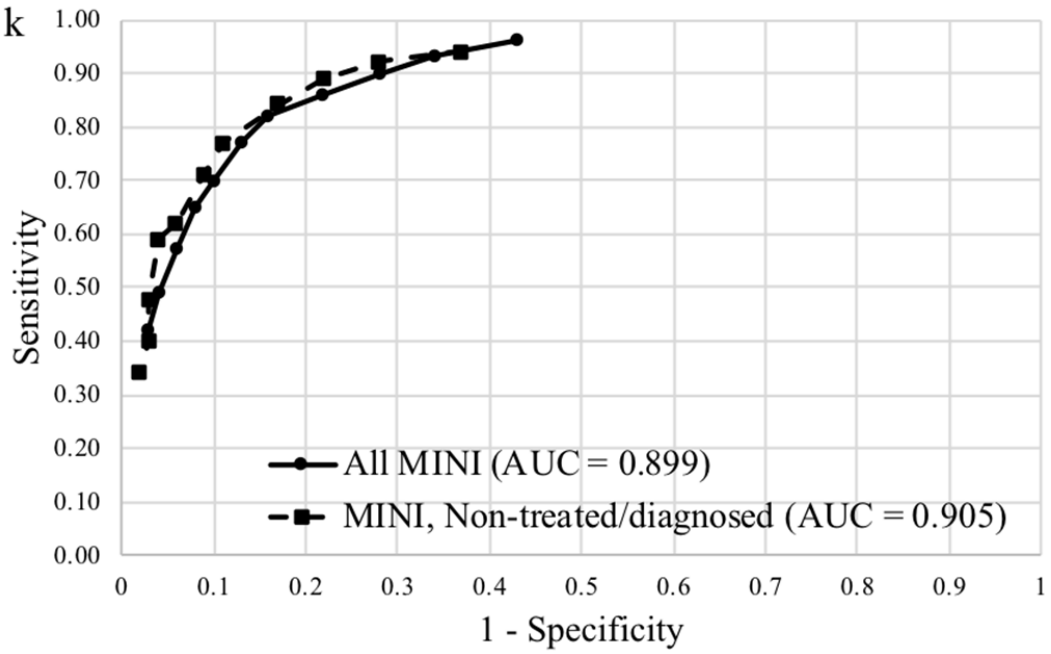
Supplementary figure D9. ROC curves comparing PHQ-9 sensitivity and specificity among participants from countries with a very high human development index, a high human development index and a low-medium human development index, among studies that used a fully structured diagnostic interview as the reference standard



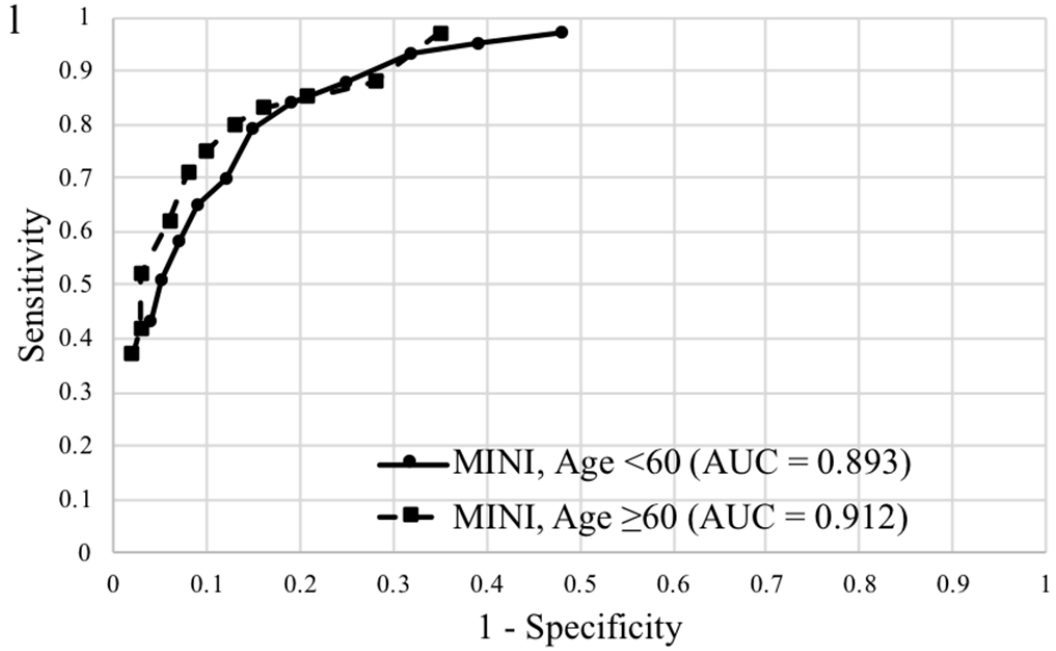
Supplementary figure D10. ROC curves comparing PHQ-9 sensitivity and specificity among participants from non-medical, primary care, inpatient speciality care and outpatient specialty care, among studies that used a fully structured diagnostic interview as the reference standard



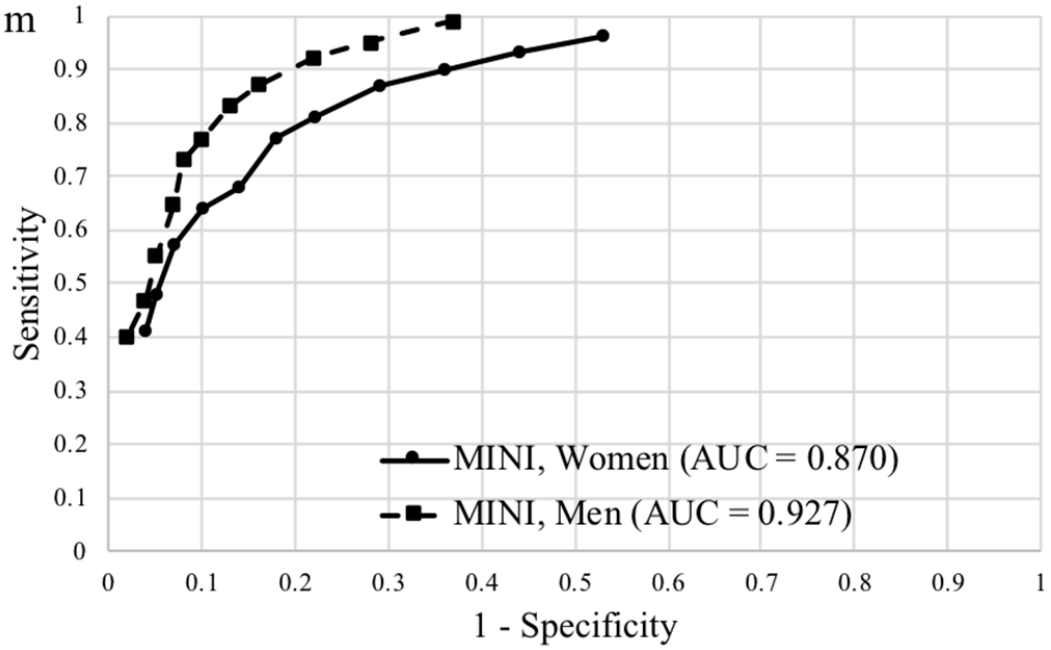
Supplementary figure D11. ROC curves comparing PHQ-9 sensitivity and specificity among all participants compared to participants not currently diagnosed or receiving treatment for a mental health problem, among studies that used the MINI as the reference standard



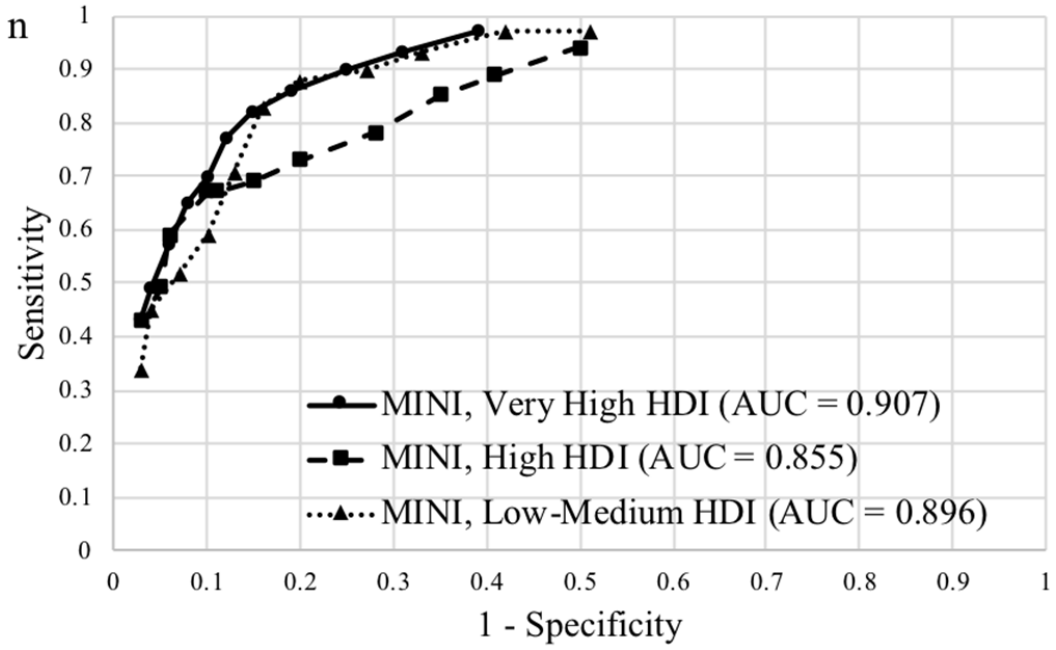
Supplementary figure D12. ROC curves comparing PHQ-9 sensitivity and specificity among among participants aged <60 compared to participants aged ≥60, among studies that used the MINI as the reference standard



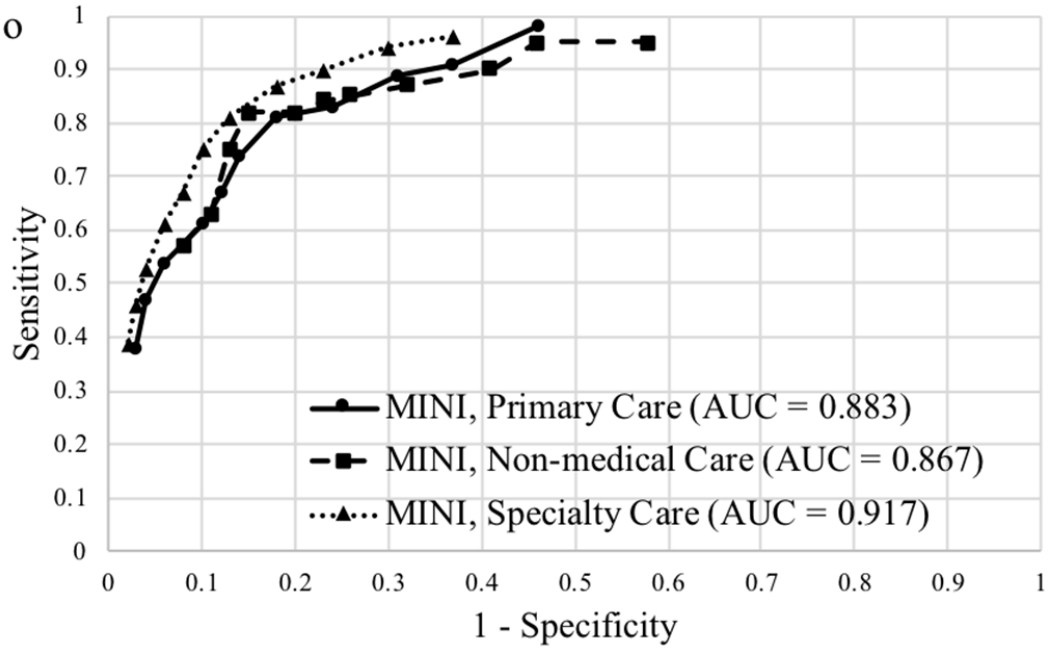
Supplementary figure D13. ROC curves comparing PHQ-9 sensitivity and specificity among among women compared to men, among studies that used the MINI as the reference standard



Supplementary figure D14. ROC curves comparing PHQ-9 sensitivity and specificity among participants from countries with a very high human development index, a high human development index and a low-medium human development index, among studies that used the MINI as the reference standard



Supplementary figure D15. ROC curves comparing PHQ-9 sensitivity and specificity among participants from non-medical, primary care, and specialty care, among studies that used the MINI as the reference standard



Supplementary table A. Reasons for exclusion for all articles excluded at full-text level (N = 113)

Reference	Reason for Exclusion
Albert NM, Moser DK, Nutter B, Pozuelo L. Are PHQ-9 and PHQ-2 Depression score cutoffs the best cutoffs for determining significant depression in Pts with HF and Mild-Moderate Symptoms? *Journal of Cardiac Failure*. 2009;**15**:S114-S114.	Major depression not assessed
Allgaier AK, Pietsch K, Fruhe B, et al. Depression in pediatric care: Is the WHO-Five Well-Being Index a valid screening instrument for children and adolescents? *General Hospital Psychiatry.* 2012;**34**:234-241.	PHQ not administered
Armstrong G, Nuken A, Samson L, et al. Quality of life, depression, anxiety and suicidal ideation among men who inject drugs in Delhi, India. *BMC Psychiatry*. 2013;**13**:151-151.	Major depression not assessed
Arroll B, Goodyear-Smith F, Kerse N, et al. The prevalence of depression among Maori patients in Auckland general practice. *Journal of Primary Health Care*. 2009;**1**:26-29.	Major depression not assessed
Berghofer A, Hartwich A, Bauer M, et al. Efficacy of a systematic depression management program in high utilizers of primary care: a randomized trial. *BMC Health Services Research*. 2012;**12**:298.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Buehler B, Kocalevent R, Berger R, et al. Treatment situation of long-term unemployed with psychological disorders. *Nervenarzt*. 2013;**84**:603-607.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Cannon DS, Tiffany ST, Coon H, et al. The PHQ-9 as a brief assessment of lifetime major depression. *Psychological Assessment.* 2007;**19**:247-251.	Major depression not assessed
Carballeira Y, Dumont P, Borgacci S, et al. Criterion validity of the French version of Patient Health Questionnaire (PHQ) in a hospital department of internal medicine. *Psychology & Psychotherapy: Theory, Research & Practice*. 2007;**80**:69-77.	No validated interview to assess major depression
Cassin S, Sockalingam S, Hawa R, et al. Psychometric properties of the Patient Health Questionnaire (PHQ-9) as a depression screening tool for bariatric surgery candidates. *Psychosomatics*. 2013;**54**:352-358.	> 2 weeks between PHQ and diagnostic interview
Chen S, Chiu H, Xu B, et al. Reliability and validity of the PHQ-9 for screening late-life depression in Chinese primary care. I*nternational Journal of Geriatric Psychiatry.* 2010;**25**:1127-1133.	> 2 weeks between PHQ and diagnostic interview
Choi Y, Mayer TG, Williams MJ, Gatchel RJ. What is the best screening test for depression in chronic spinal pain patients? *Spine Journal: Official Journal of the North American Spine Society.* 2014;**14**:1175-1182.	> 2 weeks between PHQ and diagnostic interview
Corapcioglu A, Ozer GU. Adaptation of revised Brief PHQ (Brief-PHQ-r) for diagnosis of depression, panic disorder and somatoform disorder in primary healthcare settings. *International Journal of Psychiatry in Clinical Practice*. 2004;**8**:11-18.	No validated interview to assess major depression
Creed F. The relationship between somatic symptoms, health anxiety, and outcome in medical out-patients. *Psychiatric Clinics of North America*. 2011;**34**:545-564.	PHQ not administered
Davis K, Pearlstein T, Stuart S, O'Hara M, Zlotnick C. Analysis of brief screening tools for the detection of	Sample selected for known distress,

postpartum depression: comparisons of the PRAMS 6-item instrument, PHQ-9, and structured interviews. *Archives of Women's Mental Health*. 2013;**16**:271-277.	mental health diagnosis, or psychiatric setting
de Man-van Ginkel J, Floor G, Marieke S, Eline L, Thora H. Early detection of post stroke depression: a clinimetric evaluation of the PHQ-9. *Journal of Clinical Nursing*. 2010;**19**:88-88.	Major depression not assessed
Diez-Quevedo C, Rangil T, Sanchez-Planell L, Kroenke K, Spitzer RL. Validation and utility of the Patient Health Questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. *Psychosomatic Medicine.* 2001;**63**:679-686.	No validated interview to assess major depression
Esler D, Johnston F, Thomas D, Davis B. The validity of a depression screening tool modified for use with Aboriginal and Torres Strait Islander people. *Australian & New Zealand Journal of Public Health.* 2008;**32**:317-321.	No validated interview to assess major depression
Fine TH, Contractor AA, Tamburrino M, et al. Validation of the telephone-administered PHQ-9 against the in-person administered SCID-I major depression module. *Journal of Affective Disorders*. 2013;**150**:1001-1007.	PHQ not administered
Galek A, Erbsloeh-Moeller B, Koellner V, et al. Mental disorders in patients with fibromyalgia syndrome. Screening in centres of different medical specialties. *Schmerz*. 2013;**27**:296-304.	Major depression not assessed
Gawlik S, Waldeier L, Mueller M, et al. Subclinical depressive symptoms during pregnancy and birth outcome-a pilot study in a healthy German sample. *Archives of Womens Mental Health*. 2013;**16**:93-100.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gellis ZD. Depression screening in medically ill homecare elderly. *Best Practices in Mental Health: An International Journal*. 2010;**6**:1-16.	PHQ not administered
Gibbons RD, Hooker G, Finkelman MD, et al. The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): a screening tool for depression. *Journal of Clinical Psychiatry.* 2013;**74**:669-674.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gibbons RD, Weiss DJ, Pilkonis PA, et al. Development of a computerized adaptive test for depression. *Archives of General Psychiatry.* 2012;**69**:1104-1112.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gigantesco A, Mirante N, Granchelli C, et al. Psychopathological chronic sequelae of the 2009 earthquake in L'Aquila, Italy. *Journal of Affective disorders*. 2013;**148**:265-271.	Major depression not assessed
Gilbody S, Richards D, Barkham M. Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ-9 and CORE-OM. *British Journal of General Practice.* 2007;**57**:650-652.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gold KJ, Spangenberg K, Wobil P, Schwenk TL. Depression and risk factors for depression among mothers of sick infants in Kumasi, Ghana. *International Journal of Gynaecology & Obstetrics*. 2013;**120**:228-231.	Major depression not assessed
Gothwal VK, Bagga DK, Bharani S, Sumalini R, Reddy SP. The Patient Health Questionnaire-9: Validation among patients with glaucoma. *PLoS ONE*. 2014;**9**:Art e101295-8.	Major depression not assessed
Grote NK, Katon WJ, Lohr MJ, et al. Culturally relevant treatment services for perinatal depression in socio-economically disadvantaged women: The design of the MOMCare study. *Contemporary Clinical Trials*. 2014;**39**:34-49.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Hanwella R, Ekanayake S, de Silva VA. The validity and reliability of the Sinhala translation of the Patient Health Questionnaire (PHQ-9) and PHQ-2 screener. *Depression Research and Treatment.* 2014;**2014**:768978.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Hauffa R, Rief W, Brahler E, et al. Lifetime traumatic experiences and posttraumatic stress disorder in the German population: results of a representative population survey. *Journal of Nervous & Mental Disease*. 2011;**199**:934-939.	Major depression not assessed
Hauser W, Glaesmer H, Schmutzer G, Brahler E. Widespread pain in older Germans is associated with posttraumatic stress disorder and lifetime employment status--results of a cross-sectional survey with a representative population sample. *Pain*. 2012;**153**:2466-2472.	Major depression not assessed
Hausteiner-Wiehle C, Sokollu F. Magical thinking in somatoform disorders: an exploratory study among patients with suspected allergies. *Psychopathology*. 2011;**44**:283-288.	Major depression not assessed
Holzapfel N, Muller-Tasch T, Wild B, et al. Depression profile in patients with and without chronic heart failure. *Journal of Affective Disorders*. 2008;**105**:53-62.	Major depression not assessed
Howell EA, Bodnar-Deren S, Balbierz A, et al. An intervention to reduce postpartum depressive symptoms: A randomized controlled trial. *Archives of Women's Mental Health*. 2014;**17**:57-63.	Major depression not assessed
Husain N, Creed F, Tomenson B. Depression and social stress in Pakistan. *Psychological Medicine*. 2000;**30**:395-402.	PHQ not administered
Husain N, Gater R, Tomenson B, Creed F. Comparison of the Personal Health Questionnaire and the Self Reporting Questionnaire in rural Pakistan. *JPMA - Journal of the Pakistan Medical Association.* 2006;**56**:366-370.	PHQ not administered
Husain N, Waheed W, Tomenson B, Creed F. The validation of personal health questionnaire amongst people of Pakistani family origin living in the United Kingdom. *Journal of Affective Disorders*. 2007;**97**:261-264.	PHQ not administered
Inoue T, Tanaka T, Nakagawa S. Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. *BMC Psychiatry*. 2012;**12**:73.	No validated interview to assess major depression
Jacobs SR, Jacobsen PB, Donovan K, Booth-Jones M. Utility of the Patient Health Questionnaire-9 (Phq-9) in identifying depression among hematopoietic stem cell transplant (HSCT) patients. *Annals of Behavioral Medicine*. 2007;**33**:S56-S56.	Major depression not assessed
Jeon HJ, Park JH, Shim EJ. Permissive attitude toward suicide and future intent in individuals with and without depression: results from a nationwide survey in Korea. *Journal of Nervous & Mental Disease*. 2013;**201**:286-291.	Major depression not assessed
Kamphuis MH, Stegenga BT, Zuithoff NP, et al. Does recognition of depression in primary care affect outcome? The PREDICT-NL study. *Family Practice.* 2012;**29**:16-23.	Major depression not assessed
Karekla M, Pilipenko N, Feldman J. Greek language validation of the Patient Health Questionnaire (PHQ). *Annals of Behavioral Medicine*. 2011;**41**:S20-S20.	Major depression not assessed
Kissane DW, Wein S, Love A, et al. The Demoralization Scale: a report of its development and preliminary validation. *Journal of Palliative Care*. 2004;**20**:269-276.	Major depression not assessed
Krause S, Rydall A, Hales S, Rodin G, Lo C. Initial validation of the Death and Dying Distress Scale for the assessment of death anxiety in patients with advanced cancer. *Journal of Pain and Symptom Management*. 2015;**49**:127-135.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical Care*. 2003;**41**:1284-1292.	No validated interview to assess major depression
Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*. 2001;**16**:606-613.	No validated interview to assess major depression

Lewis BA, Gjerdingen DK, Avery MD, et al. Examination of a telephone-based exercise intervention for the prevention of postpartum depression: design, methodology, and baseline data from The Healthy Mom study. *Contemporary Clinical Trials.* 2012;**33**:1150-1158.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Lewis BA, Gjerdingen DK, Avery MD, et al. A randomized trial examining a physical activity intervention for the prevention of postpartum depression: The healthy mom trial. *Mental Health and Physical Activity.* 2014;**7**:42-49.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Li C, Friedman B, Conwell Y, Fiscella K. Validity of the Patient Health Questionnaire 2 (PHQ-2) in identifying major depression in older people. *Journal of the American Geriatrics Society*. 2007;**55**:596-602.	Major depression not assessed
Lino VT, Portela MC, Camacho LA, et al. Screening for depression in low-income elderly patients at the primary care level: use of the Patient Health Questionnaire-2. *PLoS One*. 2014;**9**:e113778-e113778.	Study only administered the PHQ-2
Liu LT, Chen SL, Jin T, et al. Natural outcome and risk-prediction model of late-life depression. *Zhejiang da Xue Xue Bao Yi Xue Ban/Journal of Zhejiang University Medical Sciences.* 2012;**41**:653-658.	> 2 weeks between PHQ and diagnostic interview
Londono A, Romero P, Casas G. The association between armed conflict, violence and mental health: a cross sectional study comparing two populations in Cundinamarca department, Colombia. *Conflict & Health*. 2012;**6**:12.	Major depression not assessed
Lossnitzer N, Muller-Tasch T, Lowe B, et al. Exploring potential associations of suicidal ideation and ideas of self-harm in patients with congestive heart failure. *Depression & Anxiety*. 2009;**26**:764-768.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Lowe B, Grafe K, Kroenke K, et al. Predictors of psychiatric comorbidity in medical outpatients. *Psychosomatic Medicine*. 2003;**65**:764-770.	PHQ not administered
Lowe B, Grafe K, Quenter A, et al. The Patient Health Questionnaire D as a self-rating instrument for screening mental disorders in internal medicine and in general medicine - Preliminary validation results with 1000 outpatients. *Psychotherapie Psychosomatik Medizinische Psychologie*. 2001;**51**:109-109.	No original data
Lowe B, Grafe K, Zipfel S, et al. Detecting panic disorder in medical and psychosomatic outpatients: comparative validation of the Hospital Anxiety and Depression Scale, the Patient Health Questionnaire, a screening question, and physicians' diagnosis. *Journal of Psychosomatic Research*. 2003;**55**:515-519.	PHQ not administered
Lowe B, Kroenke K, Spitzer RL, et al. Trauma exposure and posttraumatic stress disorder in primary care patients: cross-sectional criterion standard study. *Journal of Clinical Psychiatry*. 2011;**72**:304-312.	Major depression not assessed
Mahajan S, Avasthi A, Grover S, Chawla YK. Role of baseline depressive symptoms in the development of depressive episode in patients receiving antiviral therapy for hepatitis C infection. *Journal of Psychosomatic Research*. 2014.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Maneeton B, Maneeton N, Mahathep P. Prevalence of depression and its correlations: a cross-sectional study in Thai cancer patients. *Asian Pacific Journal of Cancer Prevention: APJCP.* 2012;**13**:2039-2043.	Major depression not assessed
Mao HJ, Li HJ, Chiu H, Chan WC, Chen SL. Effectiveness of antenatal emotional self-management training program in prevention of postnatal depression in Chinese women. *Perspectives in Psychiatric Care*. 2012;**48**:218-224.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Margrove K, Mensah S, Thapar A, Kerr M. Depression screening for patients with epilepsy in a primary care setting using the Patient Health Questionnaire-2 and the Neurological Disorders Depression Inventory for Epilepsy. *Epilepsy & Behavior*. 2011;**21**:387-390.	Study only administered the PHQ-2
Mautner E, Ashida C, Greimel E, et al. Are there differences in the health outcomes of mothers in Europe and East-Asia? A cross-cultural health Survey. *Biomed Research International*. 2014;856543-856543.	Major depression not assessed

Mitchell AJ, McGlinchey JB, Young D, Chelminski I, Zimmerman M. Accuracy of specific symptoms in the diagnosis of major depressive disorder in psychiatric out-patients: data from the MIDAS project. *Psychological Medicine*. 2009;**39**:1107-1116.	PHQ not administered
Mittal D, Fortney JC, Pyne JM, Wetherell JL. Predictors of persistence of comorbid generalized anxiety disorder among veterans with major depressive disorder. *Journal of Clinical Psychiatry.* 2011;**72**:1445-1451.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Morina N, von Lersner U, Prigerson HG. War and bereavement: consequences for mental and physical distress. *PLoS ONE*. 2011;**6**:e22140.	PHQ not administered
Muller KW, Beutel ME, Wolfling K. A contribution to the clinical characterization of Internet addiction in a sample of treatment seekers: validity of assessment, severity of psychopathology and type of co-morbidity. *Comprehensive Psychiatry*. 2014;**55**:770-777.	Major depression not assessed
Mulligan L, Fear NT, Jones N, et al. Postdeployment Battlemind training for the U.K. armed forces: A cluster randomized controlled trial. *Journal of Consulting and Clinical Psychology*. 2012;**80**:331-341.	Major depression not assessed
Mussell M, Kroenke K, Spitzer RL, et al. Gastrointestinal symptoms in primary care: prevalence and association with depression and anxiety. *Journal of Psychosomatic Research*. 2008;**64**:605-612.	Major depression not assessed
Olariu E, Castro-Rodriguez JI, Alvarez P, et al. Validation of clinical symptom irt scores for diagnosis and severity assessment of common mental disorders. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation.* 2014.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Orive M, Padierna JA, Quintana JM, et al. Detecting depression in medically ill patients: Comparative accuracy of four screening questionnaires and physicians' diagnoses in Spanish population. *Journal of Psychosomatic Research.* 2010;**69**:399-406.	No validated interview to assess major depression
Osorio FL, de Carvalho AC, Crippa JA, Loureiro SR. Screening for smoking in a general hospital: scale validation, indicators of prevalence, and comorbidity. *Perspectives in Psychiatric Care*. 2013;**49**:5-12.	Major depression not assessed
Park H, Kim J, Hahm B. The Distress Thermometer and the PHQ-2 for ultra-brief screening depression of cancer patients In Korea. *Psycho-oncology.* 2013;**22**:303-304.	Study only administered the PHQ-2
Pibernik-Okanovic M, Grgurevic M, Ajdukovic D, Novak B, Begic D, Metelko Z. Screening performance of a short versus long version of the Patient Health Questionnaire-depression in outpatients with diabetes. *Diabetologia*. 2009;**52**:S392-S393.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Pilipenko N, Karekla M, Feldman J. Validation of Patient Health Questionnaire in Greek-language sample. *European Psychiatry*. 2011;**26**.	Major depression not assessed
Poutanen O, Koivisto AM, Salokangas RK. Applicability of the DEPS Depression Scale: assessing format and individual items in subgroups of patients. *Nordic Journal of Psychiatry*. 2010;**64**:384-390.	Major depression not assessed
Prescott MR, Tamburrino M, Calabrese JR, et al. Validation of lay-administered mental health assessments in a large Army National Guard cohort. *International Journal of Methods in Psychiatric Research.* 2014;**23**:109-119.	> 2 weeks between PHQ and diagnostic interview
Priyanka P, Boyle LL, Tu XM, Conwell Y. Inter-rater reliability and validity of the PHQ-9 and GAD-7 to identify depression and anxiety in older adults receiving aging services care management. *American Journal of Geriatric Psychiatry*. 2010;**18**:S113-S114.	No original data
Reck C, Stehle E, Reinig K, Mundt C. Maternity blues as a predictor of DSM-IV depression and anxiety disorders in the first three months postpartum. *Journal of Affective Disorders.* 2009;**113**:77-87.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Rentsch D, Dumont P, Borgacci S, et al. Prevalence and treatment of depression in a hospital department of internal medicine. *General Hospital Psychiatry*. 2007;**29**:25-31.	No validated interview to assess major depression
Rief W, Mewes R, Martin A, Glaesmer H, Braehler E. Are psychological features useful in classifying patients with somatic symptoms? *Psychosomatic Medicine.* 2010;**72**:648-655.	> 2 weeks between PHQ and diagnostic interview
Ringoir L, Pedersen SS, Widdershoven JW, Pop VJ. Prevalence of psychological distress in elderly hypertension patients in primary care. *Netherlands Heart Journal.* 2014;**22**:71-76.	Major depression not assessed
Rizzo R, Piccinelli M, Mazzi MA, Bellantuono C, Tansella M. The Personal Health Questionnaire: a new screening instrument for detection of ICD-10 depressive disorders in primary care. *Psychological Medicine*. 2000;**30**:831-840.	PHQ not administered
Ryan DA, Gallagher P, Wright S, Cassidy EM. Sensitivity and specificity of the Distress Thermometer and a two-item depression screen (Patient Health Questionnaire-2) with a 'help' question for psychological distress and psychiatric morbidity in patients with advanced cancer. *Psycho-oncology*. 2012;**21**:1275-1284.	PHQ not administered
Saliba D, DiFilippo S, Edelen MO, et al. Testing the PHQ-9 interview and observational versions (PHQ-9 OV) for MDS 3.0. *Journal of the American Medical Directors Association*. 2012;**13**:618-625.	PHQ not administered
Salve H, Goswami K, Nongkynrih B, Sagar R, Sreenivas V. Prevalence of psychiatric morbidity at Mobile Health Clinic in an urban community in North India. *General Hospital Psychiatry*. 2012;**34**:121-126.	PHQ not administered
Sayers SL, Farrow VA, Ross J, Oslin DW. Family problems among recently returned military veterans referred for a mental health evaluation. *Journal of Clinical Psychiatry*. 2009;**70**:163-170.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Schmitz-Hubsch T, Coudert M, Tezenas du Montcel S, et al. Depression comorbidity in spinocerebellar ataxia. *Movement Disorders*. 2011;**26**:870-876.	Major depression not assessed
Shen Q, Bergquist-Beringer S. Relationship between major depression and insulin resistance: Does it vary by gender or race/ethnicity among young adults aged 20-39 years? *Journal of Diabetes*. 2013;**5**:471-481.	Major depression not assessed
Shoukri MM, Donner A. Bivariate modeling of interobserver agreement coefficients. *Statistics in medicine.* 2009;**28**:430-440.	No original data
Smith AB, Rush R, Wright P, et al. Validation of an item bank for detecting and assessing psychological distress in cancer patients. *Psycho-oncology.* 2009;**18**:195-199.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Smith GC, McAsey P, Trauer T. Screening and monitoring in renal dialysis and transplant patients using the SF36 and Patient Health Questionnaire. *Australian and New Zealand Journal of Psychiatry*. 2000;**34**:A62-A62.	Major depression not assessed
Smith GC, McAsey P, Trauer T. Screening and monitoring in renal analysis and transplant patients using the SF36 and Patient Health Questionnaire. *Psychosomatics.* 2001;**42**:182-183.	Major depression not assessed
Smith GC, Trauer T, Kerr PG, Chadban SJ. Prospective psychosocial monitoring of living kidney donors using the Short Form-36 Health Survey: Results at 12 months. *Transplantation.* 2004;**78**:1384-1389.	No validated interview to assess major depression
Smith MV, Gotman N, Lin H, Yonkers KA. Do the PHQ-8 and the PHQ-2 accurately screen for depressive disorders in a sample of pregnant women? *General Hospital Psychiatry*. 2010;**32**:544-548.	Study only administered the PHQ-8
Sockalingam S, Blank D, Al Jarad A, et al. A comparison of depression screening instruments in hepatitis C and the impact of depression on somatic symptoms. *Psychosomatics*. 2011;**52**:433-440.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Stegenga BT, Kamphuis MH, King M, Nazareth I, Geerlings MI. The natural course and outcome of major	Major depression not assessed

depressive disorder in primary care: the PREDICT-NL study. *Social Psychiatry & Psychiatric Epidemiology*. 2012;**47**:87-95.	
Subramanian U, Perkins SM, Kim J, Ding Y, Pressler SJ. Depressive symptoms in heart failure: Validity and reliability of the PHQ-8. *Journal of General Internal Medicine*. 2008;**23**:276-276.	Major depression not assessed
Suzuki T, Shiga T, Nishimura K, Ishigooka J, Hagiwara N. PHQ-9 screening for depression in hospitalized patients with heart failure. *European Journal of Heart Failure*. 2013;S242-S242.	Major depression not assessed
Tabb KM, Gavin AR, Guo Y, et al. Views and experiences of suicidal ideation during pregnancy and the postpartum: findings from interviews with maternal care clinic patients. *Women & Health*. 2013;**53**:519-535.	Major depression not assessed
Tavakkoli M, Ferrando SJ, Rabkin J, Marks K, Talal AH. Depression and fatigue in chronic hepatitis C patients with and without HIV co-infection. *Psychosomatics*. 2013;**54**:466-471.	No validated interview to assess major depression
Thapar A, Hammerton G, Collishaw S, et al. Detecting recurrent major depressive disorder within primary care rapidly and reliably using short questionnaire measures. *British Journal of General Practice.* 2014;**64**:e31-7.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Thekkumpurath P, Walker J, Butcher I, et al. Screening for major depression in cancer outpatients: the diagnostic accuracy of the 9-item Patient Health Questionnaire. *Cancer*. 2011;**117**:218-227.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Tilli V, Suominen K, Karlsson H. The Autonomic Nervous System Questionnaire and the Brief Patient Health Questionnaire as screening instruments for panic disorder in Finnish primary care. *European Psychiatry: the Journal of the Association of European Psychiatrists.* 2013;**28**:442-447.	PHQ not administered
Tschudi-Madsen H, Kjeldsberg M, Natvig B, et al. Multiple symptoms and medically unexplained symptoms-Closely related concepts in general practitioners' evaluations. A linked doctor-patient study. *Journal of Psychosomatic Research*. 2013;**74**:186-190.	PHQ not administered
Uebelacker LA, German NM, Gaudiano BA, Miller IW. Patient Health Questionnaire depression scale as a suicide screening instrument in depressed primary care patients: a cross-sectional study. *The Primary Care Companion to CNS Disorders*. 2011;**13.**	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Ulhaq S, Symeon C, Agius M. Use of the PHQ-9 as a screening tool for post-stroke depression. *European Psychiatry*. 2010;**25**.	Major depression not assessed
Vera M, Reyes-Rabanillo ML, Huertas S, et al. Suicide ideation, plans, and attempts among general practice patients with chronic health conditions in Puerto Rico. *International Journal of General Medicine*. 2011;**4**:197-205.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Watson LC, Zimmerman S, Cohen LW, Dominik R. Practical depression screening in residential care/assisted living: five methods compared with gold standard diagnoses. *American Journal of Geriatric Psychiatry*. 2009;**17**:556-564.	PHQ not administered
Whitlow NR, Ryan GL, Stuart SP. The Patient Health Questionnaire (PHQ) is a poor psychological screening tool in in vitro fertilization (IVF) Patients. *Fertility and Sterility.* 2011;**96**:S11-S11.	Major depression not assessed
Williams LS, Brizendine EJ, Plue L, et al. Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke*. 2005;**36**:635-638.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Yeung A, Fung F, Yu SC, et al. Validation of the Patient Health Questionnaire-9 for depression screening among Chinese Americans. *Comprehensive Psychiatry*. 2008;**49**:211-217.	> 2 weeks between PHQ and diagnostic interview

Yeung A, Yu SC, Fung F, Vorono S, Fava M. Recognizing and engaging depressed Chinese Americans in treatment in a primary care setting. *International Journal of Geriatric Psychiatry.* 2006;**21**:819-823.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Zuithoff NP, Vergouwe Y, King M, et al. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. *BMC Family Practice*. 2010;**11**:98.	Major depression not assessed

Supplementary table B1. Characteristics of included primary studies

First Author, Year	Country	Recruited Population	Diagnostic Interview	Classification System	Total N	Major Depression N (%)
Semi-structured Interviews						
Amoozegar, 2017[1a]	Canada	Migraine patients	SCID	DSM-IV	203	49 (24)
Ayalon, 2010[2]	Israel	Elderly primary care patients	SCID	DSM-IV	151	6 (4)
Beraldi, 2014[3]	Germany	Cancer inpatients	SCID	DSM-IV	116	7 (6)
Bombardier, 2012[4]	USA	Inpatients with spinal cord injuries	SCID	DSM-IV	160	14 (9)
Chagas, 2013[5]	Brazil	Outpatients with Parkinson's Disease	SCID	DSM-IV	84	19 (23)
Eack, 2006[6]	USA	Women seeking psychiatric services for their children at two mental health centers	SCID	DSM-IV	48	12 (25)
Fann, 2005[7]	USA	Inpatients with traumatic brain injury	SCID	DSM-IV	134	45 (34)
Fiest, 2014[8]	Canada	Epilepsy outpatients	SCID	DSM-IV	168	23 (14)
Fischer, 2014[9]	Germany	Heart failure patients	SCID	DSM-IV	192	10 (5)
Gjerdingen, 2009[10]	USA	Mothers registering their newborns for well-child visits at medical or pediatric clinics	SCID	DSM-IV	417	19 (5)
Gräfe, 2004[11]	Germany	Medical and psychosomatic outpatients	SCID	DSM-IV	473	66 (14)
Khamseh, 2011[12]	Iran	Type 2 diabetes patients	SCID	DSM-IV	183	78 (43)
Kwan, 2012[13]	Singapore	Post-stroke inpatients undergoing rehabilitation	SCID	DSM-IV-TR	113	3 (3)
Lambert, 2015[14a]	Australia	Cancer patients	SCID	DSM-IV	147	21 (14)
Liu, 2011[15]	Taiwan	Primary care patients	SCAN	DSM-IV	1532	50 (3)
McGuire, 2013[16]	USA	Acute coronary syndrome inpatients	DISH	DSM-IV	100	9 (9)
Osório, 2009[17]	Brazil	Women in primary care	SCID	DSM-IV	177	60 (34)
Osório, 2012[18]	Brazil	Inpatients from various clinical wards	SCID	DSM-IV	86	28 (33)
Picardi, 2005[19]	Italy	Inpatients with skin diseases	SCID	DSM-IV	138	12 (9)
Richardson, 2010[20]	USA	Older adults undergoing in-home aging services care management	SCID	DSM-IV	377	95 (25)

		assessment				
Rooney, 2013[21]	UK	Adults with cerebral glioma	SCID	DSM-IV	126	14 (11)
Sidebottom, 2012[22]	USA	Pregnant women	SCID	DSM-IV	242	12 (5)
Simning, 2012[23]	USA	Older adults living in public housing	SCID	DSM-IV	190	10 (5)
Turner, Unpublished	Australia	Cardiac rehabilitation patients	SCID	DSM-IV	51	4 (8)
Turner, 2012[24]	Australia	Stroke patients	SCID	DSM-IV	72	13 (18)
Twist, 2013[25]	UK	Type 2 diabetes outpatients	SCAN	DSM-IV	360	80 (22)
Vöhringer, 2013[26]	Chile	Primary care patients	SCID	DSM-IV	190	59 (31)
Williams, 2012[27]	USA	Parkinson's Disease patients	SCID	DSM-IV	235	61 (26)
Wittkampf, 2009[28]	The Netherlands	Primary care patients at risk for depression	SCID	DSM-IV	260	45 (17)
Fully Structured Interviews						
Arroll, 2010[29]	New Zealand	Primary care patients	CIDI	DSM-IV	2523	156 (6)
Azah, 2005[30]	Malaysia	Adults attending family medicine clinics	CIDI	ICD-10	180	30 (17)
de Man-van Ginkel, 2012[31]	The Netherlands	Stroke patients	CIDI	DSM-IV	164	17 (10)
Delgadillo, 2011[32]	UK	Outpatients in drug addiction treatment	CIS-R	ICD-10	103	51 (50)
Gelaye, 2014[33]	Ethiopia	Outpatients at a general hospital	CIDI	DSM-IV	923	162 (18)
Hahn, 2006[34]	Germany	Patients with chronic illnesses from rehabilitation centers	CIDI	DSM-IV	208	17 (8)
Henkel, 2004[35]	Germany	Primary care patients	CIDI	ICD-10	430	43 (10)
Hobfoll, 2011[36]	Israel	Jewish and Palestinian residents of Jerusalem exposed to war	CIDI	DSM-IV	141	41 (29)
Kiely, 2014[37]	Australia	Community sample of adults	CIDI	ICD-10	822	33 (4)
Mohd Sidik, 2012[38]	Malaysia	Primary care patients	CIDI	DSM-IV	146	31 (21)
Patel, 2008[39]	India	Primary care patients	CIS-R	ICD-10	299	13 (4)
Pence, 2012[40]	Cameroon	HIV-infected patients	CIDI	DSM-IV	392	11 (3)
Razykov, 2013[41]	Canada	Patients with systemic sclerosis	CIDI	DSM-IV	343	13 (4)
Thombs, 2008[42]	USA	Outpatients with coronary artery disease	C-DIS	DSM-IV	1006	221 (22)
Mini International Neuropsychiatric Interviews (MINI)						
Akena, 2013[43]	Uganda	HIV/AIDS patients	MINI	DSM-IV	91	11 (12)
Cholera, 2014[44]	South Africa	Patients undergoing routine HIV counseling and testing at a primary health care clinic	MINI	DSM-IV	395	47 (12)

Hides, 2007[45]	Australia	Injection drug users accessing a needle and syringe program	MINI	DSM-IV	103	47 (46)
Hyphantis, 2011[46]	Greece	Patients with various rheumatologic disorders	MINI	DSM-IV	213	69 (32)
Hyphantis, 2014[47]	Greece	Patients with chronic illnesses presenting at the emergency department	MINI	DSM-IV	349	95 (27)
Inagaki, 2013[48]	Japan	Internal medicine outpatients	MINI	DSM-III-R	104	21 (20)
Lamers, 2008[49]	The Netherlands	Elderly primary care patients with diabetes mellitus or chronic obstructive pulmonary disease	MINI	DSM-IV	104	59 (57)
Lotrakul, 2008[50]	Thailand	Outpatients	MINI	DSM-IV	278	19 (7)
Muramatsu, 2007[51]	Japan	Primary care patients	MINI	DSM-IV	114	31 (27)
Persoons, 2001[52]	Belgium	Inpatients and patients at gastroenterological and hepatology wards	MINI	DSM-IV	173	28 (16)
Santos, 2013[53]	Brazil	General population	MINI	DSM-IV	196	25 (13)
Stafford, 2007[54]	Australia	Inpatients with coronary artery disease who had undergone surgery	MINI	DSM-IV	193	35 (18)
Sung, 2013[55]	Singapore	Primary care patients	MINI	DSM-IV	399	12 (3)
van Steenbergen-Weijenburg, 2010[56]	The Netherlands	Diabetes patients	MINI	DSM-IV	172	33 (19)
Zhang, 2013[57]	China	Type 2 diabetes patients	MINI	DSM-IV	68	17 (25)

Abbreviations: C-DIS: Computerized Diagnostic Interview Schedule; CIDI: Composite International Diagnostic Interview; CIS-R: Clinical Interview Schedule Revised; DISH: Depression Interview and Structured Hamilton; DSM: Diagnostic and Statistical Manual of Mental Disorders; ICD: International Classification of Diseases; MINI: Mini Neurospsychiatric Diagnostic Interview; PHQ-9: Patient Health Questionnaire-9; SCAN: Schedules for Clinical Assessment in Neuropsychiatry; SCID: Structured Clinical Interview for DSM Disorders; UK: United Kingdom; USA: United States of America.

[a]Was unpublished at the time of electronic database search

Supplementary table B2. Characteristics of eligible primary studies not included in the present study

First Author, Year	Country	Recruited Population	Diagnostic Interview	Classification System	Total N	Major Depression N (%)	Could study have been added as a published dataset? (Reason)
Semi-structured Interviews							
Becker, 2002[58]	Saudi Arabia	Primary care patients	SCID	DSM-III-R	173	NR	No (Primary study did not report accuracy results for any PHQ-9 cutoff)
Chen, 2013[59]	China	Primary care populations	SCID	DSM-IV	280	NR[a]	No (Primary study did not report the number of participants with major depression)
Chen, 2012[60]	China	Adults over 60 in primary care	SCID	DSM-IV	262	97 (37)	No (Primary study did not report accuracy results for any PHQ-9 cutoff)
Lai, 2010[61]	Hong Kong	Men with postpartum wives	SCID	DSM-IV	551	8 (1)	No (Pubished data ineligible: some participants had time intervals between PHQ-9 adminiatration and diagnostic interview that were greater than 2 weeks)
Navinés, 2012[62]	Spain	Chronic hepatitis C patients	SCID	DSM-IV	104	21 (20)	Yes (Published accuracy results for PHQ-9 cutoff 9)
Phelan, 2010[63]	USA	Elderly primary care patients	SCID	DSM-IV	69	8 (12)	Yes (Published accuracy results for PHQ-9 cutoffs 8-12)
Thompson, 2011[64]	USA	Parkinson's patients	SCID	DSM-IV	214	30 (14)	No (Primary study did not report accuracy results for any PHQ-9 cutoff)
Watnick, 2005[65]	USA	Long term dialysis patients	SCID	DSM-IV	62	12 (19)	No (Published data ineligible: reported accuracy estimates were not for major depression, they were for a broader definition of depression)
Fully Structured Interviews							
Al-Ghafri, 2014[66]	Oman	Medical trainees	CIDI	NR	131	NR[a]	No (Primary study did not report sample size or number of participants with major depression)
Haddad, 2013[67]	UK	Coronary heart disease patients	CIS-R	ICD-10	730	32 (4)	Yes (Published accuracy results for PHQ-9 cutoffs 0-24)

Mini International Neuropsychiatric Interviews (MINI)							
Persoons, 2003[68]	Belgium	Otorhinolaryngology outpatients	MINI	DSM-IV	97	16 (16)	No (Primary study did not report accuracy results for any PHQ-9 cutoff)
Rathore, 2014[69]	USA	Adults with epilepsy	MINI	DSM-IV	172	33 (19)	Yes (Published accuracy results for PHQ-9 cutoffs 10-15)
Scott, 2011[70]	USA	Chronic hepatitis C patients	MINI	DSM-IV and ICD-10	30	NR[a]	No (Primary study did not report the number of participants with major depression)
Wang, 2014[71]	China	General population	MINI	DSM-IV	1045	28 (3)	No (Published data ineligible: some participants were under the age of 18)

Abbreviations: CIDI: Composite International Diagnostic Interview; CIS-R: Clinical Interview Schedule Revised; DSM: Diagnostic and Statistical Manual of Mental Disorders; ICD: International Classification of Diseases; MINI: Mini International Neuropsychiatric Interview; NR: Not Reported; PHQ-9: Patient Health Questionnaire-9; SCID: Structured Clinical Interview for DSM Disorders; UK: United Kingdom; USA: United States of America.

[a]Reported numbers implausible

Supplementary table C. Estimates of heterogeneity at PHQ-9 cutoff score of 10

Participant Subgroup	Semi-structured Diagnostic Interviews				Fully Structured Diagnostic Interviews				Mini International Neuropsychiatric Interviews			
	R^a		τ^2		R^a		τ^2		R^a		τ^2	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
All participants	2.33	2.99	0.78	0.33	3.64	6.42	0.76	0.68	2.20	2.68	0.50	0.31
Participants not currently diagnosed or receiving treatment for a mental health problem	2.58	2.95	1.49	0.50	3.23	6.84	0.71	0.91	1.60	1.53	0.20	0.13
Age <60	2.11	2.78	0.93	0.34	3.31	5.74	0.84	0.68	1.68	2.37	0.40	0.27
Age ≥60	2.78	1.90	0.98	0.24	1.56	3.60	0.04	0.59	1.93	1.84	0.35	0.33
Women	2.48	2.83	1.35	0.43	2.29	6.06	0.41	0.99	1.76	2.60	0.40	0.45
Men	1.70	1.73	0.45	0.16	3.13	3.78	0.97	0.50	1.62	2.45	0.53	0.62
Very high country human development index	1.96	2.64	0.48	0.23	3.59	6.94	0.67	0.71	2.69	3.05	0.71	0.50
High country human development index	7.07	4.44	7.72	1.38	1.97	1.72	0.38	0.16	1.00	1.00	0.00	0.00
Low-medium country human development index	--	--	--	--	2.10	5.23	0.07	0.40	1.00	1.00	0.00	0.00
Non-medical care	1.00	1.00	0.00	0.00	1.47	2.67	0.12	0.14	1.41	2.47	0.20	0.27
Primary care	2.07	5.34	0.62	0.92	1.87	3.74	0.18	0.18	2.38	1.86	0.61	0.09
Inpatient specialty care[b]	1.24	1.21	0.11	0.03	1.33	2.75	0.30	0.17	--	--	--	--
Outpatient specialty care[b]	1.86	2.26	0.30	0.19	5.67	8.54	1.29	1.11	2.24	2.39	0.49	0.33

[a]R is the ratio of the estimated standard deviation of the pooled sensitivity (or specificity) from the random-effects model to the estimated standard deviation of the pooled sensitivity (or specificity) from the corresponding fixed-effects model

[b]Among studies that used the MINI as the reference standard, only 1 study included participants from an inpatient specialty care setting. These participants were combined with participants from outpatient specialty care settings for all subgroup analyses

Supplementary table D1. Coefficients and p-values for one-stage meta-regressions assessing interactions between reference standard category and logit(sensitivity) and logit(specificity)

Cutoff	5		6		7		8		9		10		11		12		13		14		15	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
d0[a]	-0.215	0.083	-0.553	<0.001	-0.850	<0.001	-1.123	<0.001	-1.437	<0.001	-1.793	<0.001	-2.083	<0.001	-2.361	<0.001	-2.665	<0.001	-2.997	<0.001	-3.236	<0.001
d0fully	0.058	0.786	0.072	0.737	0.045	0.830	0.010	0.961	0.061	0.769	0.133	0.564	0.153	0.520	0.199	0.413	0.106	0.708	0.102	0.738	0.038	0.905
d0mini	-0.096	0.651	-0.098	0.642	-0.122	0.560	-0.179	0.387	-0.206	0.322	-0.127	0.582	-0.111	0.641	-0.112	0.646	-0.146	0.608	-0.145	0.638	-0.304	0.345
d1[b]	3.910	<0.001	3.741	<0.001	3.493	<0.001	2.920	<0.001	2.374	<0.001	2.010	<0.001	1.666	<0.001	1.307	<0.001	0.883	<0.001	0.624	<0.001	0.263	0.056
d1fully	-1.198	0.014	-1.327	0.011	-1.586	0.001	-1.345	0.001	-1.296	<0.001	-1.145	<0.001	-1.171	<0.001	-1.026	<0.001	-0.902	<0.001	-0.878	<0.001	-0.857	<0.001
d1mini	-0.453	0.364	-1.116	0.028	-1.275	0.009	-1.083	0.008	-0.898	0.007	-0.814	0.008	-0.846	0.002	-0.711	0.005	-0.597	0.009	-0.678	0.003	-0.615	0.006

[a]d0 corresponds to the model coefficient for logit(specificity)
[b]d1 corresponds to the model coefficient for logit(sensitivity)

Supplementary table D2. Coefficients and p-values for one-stage meta-regressions assessing interactions between subgrouping variables and logit(sensitivity) and logit(specificity), among participants administered a semi-structured diagnostic interview

Cutoff	5		6		7		8		9		10		11		12		13		14		15	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
d0[a]	0.552	0.007	0.195	0.333	0.002	0.992	-0.399	0.050	-0.792	<0.001	-1.268	<0.001	-1.525	<0.001	-1.891	<0.001	-2.333	<0.001	-2.790	<0.001	-3.131	<0.001
d0sex	-0.342	<0.001	-0.283	<0.001	-0.285	<0.001	-0.298	<0.001	-0.319	<0.001	-0.308	<0.001	-0.358	<0.001	-0.378	<0.001	-0.337	0.003	-0.342	0.007	-0.380	0.007
d0age	-0.015	<0.001	-0.015	<0.001	-0.016	<0.001	-0.013	<0.001	-0.010	<0.001	-0.006	0.030	-0.006	0.033	-0.005	0.135	-0.002	0.423	<0.001	0.939	0.001	0.732
d0hdi.h	0.477	0.183	0.487	0.158	0.447	0.190	0.469	0.152	0.371	<0.001	0.172	0.644	0.192	0.615	0.126	0.747	0.220	0.624	0.347	0.482	0.491	0.340
d0nonmed	0.734	0.131	0.614	0.188	0.517	0.262	0.369	0.404	0.109	0.781	-0.250	0.618	-0.381	0.461	-0.408	0.435	-0.461	0.443	-0.617	0.356	-0.495	0.482
d0inpt	0.459	0.102	0.396	0.152	0.242	0.373	0.250	0.343	0.193	<0.001	0.218	0.460	0.276	0.366	0.266	0.396	0.393	0.269	0.414	0.291	0.486	0.245
d0outpt	-0.093	0.583	-0.106	0.526	-0.170	0.316	-0.197	0.256	-0.158	<0.001	-0.332	0.095	-0.305	0.149	-0.232	0.292	-0.328	0.181	-0.335	0.215	-0.281	0.341
d1[b]	5.932	<0.001	6.100	<0.001	6.140	<0.001	4.861	<0.001	4.292	<0.001	3.752	<0.001	3.195	<0.001	2.417	<0.001	1.845	<0.001	1.671	<0.001	1.257	<0.001
d1sex	0.093	0.821	-0.086	0.813	-0.042	0.897	0.120	0.667	0.327	<0.001	-0.142	0.485	-0.226	0.226	-0.210	0.218	-0.199	0.209	-0.172	0.264	-0.180	0.228
d1age	-0.028	0.042	-0.029	0.023	-0.032	0.006	-0.023	0.044	-0.030	<0.001	-0.025	<0.001	-0.023	0.001	-0.016	0.008	-0.011	0.038	-0.013	0.015	-0.011	0.026
d1hdi.h	1.025	0.109	0.660	0.235	0.649	0.237	0.858	0.132	0.881	<0.001	1.128	0.001	1.089	0.001	0.860	0.004	0.497	0.056	0.503	0.066	0.581	0.027
d1nonmed	-0.647	0.996	-0.993	0.679	-1.489	0.538	-1.247	0.997	-0.515	<0.001	-0.396	0.726	-0.356	0.311	-0.382	0.302	-0.481	0.883	-0.322	0.857	-0.692	0.566
d1inpt	16.185	0.996	4.883	0.679	3.490	0.538	-0.003	0.997	0.273	<0.001	0.180	0.726	0.477	0.311	0.405	0.302	0.047	0.883	-0.057	0.857	-0.170	0.566
d1outpt	-1.980	0.001	-2.367	<0.001	-2.680	<0.001	-2.112	<0.001	-1.696	<0.001	-1.598	<0.001	-1.334	<0.001	-1.137	<0.001	-1.050	<0.001	-1.015	<0.001	-0.897	<0.001

[a]d0 corresponds to the model coefficient for logit(specificity)
[b]d1 corresponds to the model coefficient for logit(sensitivity)

Supplementary table D3. Coefficients and p-values for one-stage meta-regressions assessing interactions between subgrouping variables and logit(sensitivity) and logit(specificity), among participants administered a fully structured diagnostic interview

Cutoff	5		6		7		8		9		10		11		12		13		14		15	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
d0[a]	1.001	0.006	0.560	0.119	0.277	0.423	0.024	0.944	-0.255	0.481	-0.561	0.135	-0.867	0.017	-1.007	0.005	-1.156	0.005	-1.634	<0.001	-1.801	<0.001
d0sex	-0.256	<0.001	-0.270	<0.001	-0.251	<0.001	-0.258	<0.001	-0.284	<0.001	-0.220	0.006	-0.298	0.001	-0.336	<0.001	-0.315	0.003	-0.263	0.029	-0.476	0.001
d0age	-0.019	<0.001	-0.019	<0.001	-0.019	<0.001	-0.020	<0.001	-0.019	<0.001	-0.020	<0.001	-0.019	<0.001	-0.020	<0.001	-0.022	<0.001	-0.019	<0.001	-0.017	<0.001
d0hdi.h	-0.629	0.236	-0.743	0.156	-0.806	0.109	-0.993	0.046	-1.218	0.020	-1.214	0.026	-1.240	0.018	-1.370	0.007	-2.208	<0.001	-2.548	<0.001	-2.743	<0.001
d0hdi.lm	-0.923	0.027	-0.935	0.023	-0.843	0.033	-0.893	0.022	-0.888	0.029	-0.939	0.027	-0.886	0.029	-0.921	0.019	-0.853	0.064	-0.820	0.077	-0.925	0.039
d0nonmed	-0.719	0.180	-0.524	0.322	-0.501	0.324	-0.514	0.307	-0.482	0.360	-0.461	0.399	-0.477	0.363	-0.430	0.397	-0.453	0.447	-0.466	0.436	-0.557	0.334
d0inpt	1.061	0.049	1.252	0.018	1.165	0.022	1.121	0.025	1.129	0.031	1.098	0.043	1.149	0.026	1.128	0.024	1.011	0.084	1.070	0.067	0.850	0.129
d0outpt	0.295	0.477	0.425	0.299	0.335	0.393	0.442	0.253	0.445	0.270	0.545	0.193	0.502	0.208	0.494	0.201	0.361	0.426	0.344	0.449	0.164	0.705
d1[b]	4.323	<0.001	3.752	<0.001	3.125	<0.001	2.896	<0.001	2.505	<0.001	2.388	<0.001	2.015	<0.001	1.671	<0.001	1.326	0.002	1.017	0.026	0.607	0.151
d1sex	-0.068	0.785	0.089	0.684	-0.129	0.501	-0.078	0.664	-0.037	0.828	0.097	0.557	0.025	0.878	0.005	0.974	0.042	0.794	-0.041	0.804	-0.083	0.624
d1age	-0.020	0.032	-0.018	0.025	-0.019	0.007	-0.021	0.002	-0.019	0.002	-0.021	0.001	-0.020	0.001	-0.019	0.001	-0.018	0.002	-0.017	0.004	-0.015	0.014
d1hdi.h	-0.931	0.397	-0.707	0.511	-0.072	0.937	0.009	0.993	-0.852	0.152	-1.065	0.049	-1.014	0.033	-0.845	0.066	-0.847	0.110	-0.813	0.149	-0.856	0.086
d1hdi.lm	-1.401	0.117	-1.687	0.052	-2.032	0.006	-1.685	0.023	-1.350	0.008	-1.313	0.005	-1.118	0.007	-1.250	0.002	-1.119	0.016	-1.119	0.022	-1.141	0.011
d1nonmed	-1.122	0.302	-0.988	0.349	-0.819	0.347	-0.810	0.370	-1.048	0.080	-1.088	0.046	-1.111	0.022	-1.083	0.022	-1.090	0.045	-0.976	0.090	-0.987	0.057
d1inpt	-1.122	0.686	-0.988	0.473	-0.819	0.373	-0.810	0.440	-1.048	0.367	-1.088	0.255	-1.111	0.396	-1.083	0.181	-1.090	0.478	-0.976	0.347	-0.987	0.290
d1outpt	-0.508	0.573	-0.302	0.724	0.233	0.745	-0.007	0.993	-0.295	0.544	-0.445	0.313	-0.482	0.215	-0.320	0.395	-0.375	0.389	-0.308	0.505	-0.295	0.480

[a]d0 corresponds to the model coefficient for logit(specificity)
[b]d1 corresponds to the model coefficient for logit(sensitivity)

Supplementary table D4. Coefficients and p-values for one-stage meta-regressions assessing interactions between subgrouping variables and logit(sensitivity) and logit(specificity), among participants administered the MINI

Cutoff	5		6		7		8		9		10		11		12		13		14		15	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
d0[a]	0.381	0.138	0.112	0.671	0.032	0.907	-0.099	0.746	-0.578	0.071	-0.804	0.039	-1.093	0.008	-1.222	0.006	-1.366	0.008	-1.787	0.002	-1.948	0.002
d0sex	-0.610	<0.001	-0.615	<0.001	-0.609	<0.001	-0.716	<0.001	-0.586	<0.001	-0.619	<0.001	-0.613	<0.001	-0.531	0.001	-0.523	0.003	-0.406	0.038	-0.554	0.017
d0age	-0.010	0.001	-0.010	0.001	-0.013	<0.001	-0.016	<0.001	-0.013	0.001	-0.015	0.001	-0.013	0.011	-0.017	0.002	-0.020	0.002	-0.021	0.004	-0.024	0.005
d0hdi.h	0.270	0.377	0.262	0.402	0.309	0.336	0.324	0.369	0.097	0.796	0.134	0.773	-0.066	0.892	0.119	0.818	-0.074	0.904	0.054	0.934	-0.159	0.831
d0hdi.lm	0.414	0.219	0.308	0.371	0.149	0.673	0.056	0.888	-0.031	0.940	-0.080	0.878	0.005	0.993	-0.036	0.950	-0.253	0.716	-0.387	0.610	-0.362	0.664
d0nonmed	0.792	0.042	0.607	0.124	0.704	0.080	0.738	0.101	0.732	0.113	0.898	0.117	0.935	0.114	0.895	0.158	1.152	0.121	1.289	0.103	1.376	0.117
d0inpt	0.301	0.440	0.214	0.593	0.107	0.802	0.065	0.895	0.176	0.732	0.477	0.470	0.195	0.779	0.224	0.765	0.365	0.671	-0.311	0.757	-0.492	0.668
d0outpt	-0.192	0.437	-0.263	0.300	-0.340	0.194	-0.315	0.287	-0.398	0.194	-0.429	0.265	-0.553	0.167	-0.550	0.201	-0.608	0.234	-0.420	0.444	-0.426	0.486
d1[b]	3.213	0.001	1.074	0.213	0.670	0.430	0.492	0.506	0.358	0.608	0.489	0.378	-0.087	0.867	0.271	0.531	0.473	0.177	0.221	0.501	-0.196	0.568
d1sex	-0.115	0.787	1.613	<0.001	1.317	<0.001	0.947	<0.001	0.698	0.002	0.514	0.012	0.620	0.001	0.440	0.014	0.281	0.101	0.239	0.158	0.156	0.370
d1age	0.023	0.055	0.008	0.481	0.014	0.221	0.015	0.169	0.016	0.101	0.007	0.414	0.012	0.119	<0.001	0.942	-0.010	0.077	-0.010	0.036	-0.007	0.175
d1hdi.h	-1.087	0.161	-0.157	0.867	-0.254	0.785	-0.543	0.485	-0.614	0.415	-0.506	0.370	-0.329	0.535	-0.029	0.943	0.071	0.830	0.050	0.870	-0.012	0.970
d1hdi.lm	-0.085	0.933	2.102	0.111	1.695	0.162	1.402	0.167	1.481	0.128	0.921	0.185	0.541	0.377	0.092	0.842	-0.071	0.837	-0.244	0.445	-0.349	0.329
d1nonmed	-1.282	0.204	1.053	0.389	0.424	0.718	0.454	0.644	0.499	0.600	0.824	0.242	0.822	0.211	1.127	0.024	1.128	0.003	0.870	0.007	0.921	0.007
d1inpt	-1.282	0.996	1.053	0.973	0.424	0.338	0.454	0.267	0.499	0.108	0.824	0.199	0.822	0.327	1.127	0.174	1.128	0.037	0.870	0.046	0.921	0.100
d1outpt	-1.585	0.029	0.385	0.626	0.165	0.836	0.049	0.941	-0.323	0.620	-0.101	0.832	-0.308	0.488	-0.173	0.593	-0.007	0.977	0.002	0.992	0.001	0.995

[a]d0 corresponds to the model coefficient for logit(specificity)
[b]d1 corresponds to the model coefficient for logit(sensitivity)

Supplementary table D5. Coefficients and p-values for one-stage meta-regressions assessing interactions between age and logit(sensitivity) and logit(specificity), among participants administered a semi-structured diagnostic interview

Cutoff	5		6		7		8		9		10		11		12		13		14		15	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
d0[a]	0.012	<0.001	-0.305	0.009	-0.601	<0.001	-0.895	<0.001	-1.221	<0.001	-1.598	<0.001	-1.862	<0.001	-2.121	<0.001	-2.440	<0.001	-2.776	<0.001	-3.041	<0.001
d0age	-0.552	<0.001	-0.595	<0.001	-0.614	<0.001	-0.572	<0.001	-0.550	<0.001	-0.497	<0.001	-0.569	<0.001	-0.624	<0.001	-0.543	<0.001	-0.531	<0.001	-0.473	0.003
d1[b]	4.056	<0.001	3.961	<0.001	3.917	<0.001	3.012	<0.001	2.418	<0.001	1.987	<0.001	1.687	<0.001	1.286	<0.001	0.880	<0.001	0.649	<0.001	0.306	0.034
d1age	-0.186	<0.001	-0.179	0.691	-0.359	0.386	-0.103	0.760	-0.085	0.765	0.274	0.313	0.089	0.711	0.107	0.629	-0.003	0.988	-0.073	0.701	-0.117	0.536

[a]d0 corresponds to the model coefficient for logit(specificity)
[b]d1 corresponds to the model coefficient for logit(sensitivity)

Supplementary table D6. Coefficients and p-values for one-stage meta-regressions assessing interactions between age and logit(sensitivity) and logit(specificity), among participants administered a fully structured diagnostic interview

Cutoff	5		6		7		8		9		10		11		12		13		14		15	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
d0[a]	<0.001	0.999	-0.336	0.140	-0.654	0.003	-0.952	<0.001	-1.223	<0.001	-1.503	<0.001	-1.792	<0.001	-2.005	<0.001	-2.386	<0.001	-2.762	<0.001	-3.063	<0.001
d0age	-0.586	<0.001	-0.545	<0.001	-0.597	<0.001	-0.657	<0.001	-0.631	<0.001	-0.681	<0.001	-0.600	<0.001	-0.712	<0.001	-0.854	<0.001	-0.721	<0.001	-0.686	<0.001
d1[b]	2.967	<0.001	2.626	<0.001	2.036	<0.001	1.734	<0.001	1.190	<0.001	1.006	<0.001	0.623	0.007	0.406	0.090	0.100	0.672	-0.149	0.539	-0.526	0.025
d1age	-0.945	0.002	-0.933	0.001	-0.664	0.004	-0.741	0.001	-0.622	0.002	-0.723	<0.001	-0.723	<0.001	-0.690	<0.001	-0.743	<0.001	-0.697	0.001	-0.623	0.004

[a]d0 corresponds to the model coefficient for logit(specificity)
[b]d1 corresponds to the model coefficient for logit(sensitivity)

Supplementary table D7. Coefficients and p-values for one-stage meta-regressions assessing interactions between age and logit(sensitivity) and logit(specificity), among participants administered the MINI

Cutoff	5		6		7		8		9		10		11		12		13		14		15	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
d0[a]	-0.229	0.138	-0.551	<0.001	-0.819	<0.001	-1.138	<0.001	-1.498	<0.001	-1.764	<0.001	-2.054	<0.001	-2.338	<0.001	-2.635	<0.001	-2.995	<0.001	-3.382	<0.001
d0age	-0.228	0.047	-0.286	0.013	-0.423	0.001	-0.451	0.001	-0.396	0.008	-0.427	0.011	-0.379	0.042	-0.378	0.068	-0.494	0.043	-0.420	0.115	-0.485	0.114
d1[b]	3.217	<0.001	2.751	<0.001	2.207	<0.001	1.713	<0.001	1.344	<0.001	1.060	<0.001	0.599	0.011	0.511	0.013	0.283	0.129	0.039	0.801	-0.290	0.072
d1age	0.473	0.355	-0.423	0.266	-0.143	0.666	0.222	0.476	0.435	0.122	0.337	0.193	0.575	0.018	0.187	0.397	-0.038	0.859	-0.302	0.131	-0.206	0.316

[a]d0 corresponds to the model coefficient for logit(specificity)
[b]d1 corresponds to the model coefficient for logit(sensitivity)

Supplementary table E1. Comparison of PHQ-9 sensitivity and specificity estimates at cutoff 10 among all participants, among participants not currently diagnosed or receiving treatment for a mental health problem, and among participant subgroups based on age, sex, human development index, and care setting

Participant Subgroup	Semi-structured Diagnostic Interviews				Fully Structured Diagnostic Interviews				Mini International Neuropsychiatric Interviews			
	Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
All participants	0.88	(0.83, 0.92)	0.85	(0.82, 0.88)	0.70	(0.59, 0.80)	0.84	(0.77, 0.89)	0.77	(0.68, 0.83)	0.87	(0.83, 0.90)
Participants not currently diagnosed or receiving treatment for a mental health problem	0.88	(0.77, 0.94)	0.89	(0.85, 0.92)	0.76	(0.59, 0.87)	0.88	(0.76, 0.94)	0.71	(0.59, 0.81)	0.91	(0.88, 0.94)
Age <60	0.87	(0.81, 0.92)	0.84	(0.80, 0.87)	0.72	(0.60, 0.82)	0.82	(0.75, 0.88)	0.79	(0.70, 0.85)	0.85	(0.80, 0.88)
Age ≥60	0.91	(0.82, 0.96)	0.88	(0.85, 0.91)	0.55	(0.44, 0.65)	0.86	(0.78, 0.91)	0.75	(0.64, 0.84)	0.90	(0.86, 0.94)
Women	0.91	(0.84, 0.95)	0.84	(0.79, 0.87)	0.67	(0.57, 0.76)	0.82	(0.73, 0.89)	0.77	(0.68, 0.84)	0.82	(0.76, 0.87)
Men	0.86	(0.79, 0.90)	0.87	(0.85, 0.89)	0.72	(0.57, 0.83)	0.86	(0.80, 0.90)	0.77	(0.66, 0.85)	0.90	(0.85, 0.94)
Very high country human development index	0.86	(0.80, 0.90)	0.86	(0.83, 0.88)	0.78	(0.65, 0.87)	0.80	(0.70, 0.88)	0.77	(0.65, 0.86)	0.88	(0.82, 0.92)
High country human development index	0.99	(0.64, 1.00)	0.86	(0.65, 0.95)	0.63	(0.38, 0.83)	0.92	(0.84, 0.96)	0.69	(0.56, 0.79)	0.85	(0.81, 0.88)
Low-medium country human development index	--	--	--	--	0.47	(0.32, 0.62)	0.88	(0.77, 0.94)	0.83	(0.71, 0.90)	0.84	(0.81, 0.87)
Non-medical care	0.82	(0.73, 0.88)	0.88	(0.85, 0.91)	0.61	(0.44, 0.75)	0.88	(0.80, 0.93)	0.84	(0.68, 0.93)	0.77	(0.60, 0.88)
Primary care	0.94	(0.88, 0.97)	0.88	(0.79, 0.93)	0.71	(0.60, 0.80)	0.88	(0.84, 0.92)	0.74	(0.56, 0.86)	0.86	(0.82, 0.89)
Inpatient specialty care[a]	0.92	(0.84, 0.96)	0.81	(0.78, 0.85)	0.89	(0.68, 0.97)	0.69	(0.54, 0.80)	--	--	--	--
Outpatient specialty care[a]	0.77	(0.67, 0.84)	0.84	(0.80, 0.88)	0.63	(0.38, 0.83)	0.80	(0.62, 0.91)	0.75	(0.63, 0.84)	0.90	(0.85, 0.93)

Abbreviations: CI: confidence interval

[a]Among studies that used the MINI as the reference standard, only 1 study included participants from an inpatient specialty care setting. These participants were combined with participants from outpatient specialty care settings for all subgroup analyses

Supplementary table E2. Comparison of PHQ-9 sensitivity and specificity estimates among participants not currently diagnosed or receiving treatment for a mental health problem compared to all participants, among participants administered a semi-structured diagnostic interview

	All participants[a]				Participants not currently diagnosed or receiving treatment for a mental health problem[b]				Difference across groups[c] (All participants – participants not currently diagnosed or receiving treatment for a mental health problem)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.98	(0.96, 0.99)	0.55	(0.49, 0.60)	1.00	(0.75, 1.00)	0.58	(0.51, 0.65)	-0.02	(-0.03, 0.01)	-0.03	(-0.10, 0.02)
6	0.98	(0.95, 0.99)	0.63	(0.58, 0.67)	0.99	(0.92, 1.00)	0.67	(0.60, 0.73)	-0.01	(-0.03, 0.03)	-0.04	(-0.10, 0.02)
7	0.98	(0.94, 0.99)	0.69	(0.65, 0.74)	0.98	(0.89, 1.00)	0.73	(0.67, 0.79)	0.00	(-0.03, 0.06)	-0.04	(-0.09, 0.01)
8	0.95	(0.91, 0.97)	0.75	(0.71, 0.79)	0.95	(0.88, 0.98)	0.79	(0.74, 0.84)	0.00	(-0.05, 0.06)	-0.04	(-0.09, 0.00)
9	0.91	(0.87, 0.94)	0.8	(0.77, 0.83)	0.91	(0.84, 0.95)	0.84	(0.80, 0.88)	0.00	(-0.05, 0.08)	-0.04	(-0.07, -0.00)
10	0.88	(0.83, 0.92)	0.85	(0.82, 0.88)	0.88	(0.77, 0.94)	0.89	(0.85, 0.92)	0.00	(-0.06, 0.12)	-0.04	(-0.07, -0.00)
11	0.84	(0.78, 0.89)	0.89	(0.86, 0.91)	0.82	(0.71, 0.90)	0.91	(0.88, 0.94)	0.02	(-0.07, 0.15)	-0.02	(-0.06, 0.00)
12	0.79	(0.73, 0.83)	0.91	(0.89, 0.93)	0.73	(0.63, 0.81)	0.94	(0.91, 0.95)	0.06	(-0.04, 0.19)	-0.03	(-0.05, 0.00)
13	0.70	(0.65, 0.75)	0.93	(0.91, 0.95)	0.66	(0.57, 0.73)	0.95	(0.93, 0.97)	0.04	(-0.04, 0.16)	-0.02	(-0.04, 0.00)
14	0.64	(0.58, 0.70)	0.95	(0.93, 0.96)	0.59	(0.49, 0.68)	0.97	(0.95, 0.98)	0.05	(-0.04, 0.20)	-0.02	(-0.03, -0.00)
15	0.56	(0.50, 0.62)	0.96	(0.95, 0.97)	0.50	(0.39, 0.60)	0.97	(0.96, 0.98)	0.06	(-0.05, 0.22)	-0.01	(-0.03, 0.00)

[a]N Studies = 29; N Participants = 6,725; N major depression = 924
[b]N Studies = 20; N Participants = 2,942; N major depression = 421
[c]20 bootstrap iterations (2%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E3. Comparison of PHQ-9 sensitivity and specificity estimates among participants aged <60 compared to ≥60, among participants administered a semi-structured diagnostic interview

	Age <60[a]				Age ≥60[b]				Difference across groups[c] (Age <60 – Age ≥60)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.98	(0.96, 0.99)	0.52	(0.46, 0.57)	0.98	(0.91, 1.00)	0.59	(0.53, 0.65)	0.00	(-0.02, 0.05)	-0.07	(-0.15, 0.01)
6	0.98	(0.95, 0.99)	0.59	(0.54, 0.65)	0.98	(0.90, 1.00)	0.68	(0.62, 0.73)	0.00	(-0.03, 0.05)	-0.09	(-0.16, 0.01)
7	0.98	(0.93, 0.99)	0.66	(0.61, 0.71)	0.97	(0.89, 0.99)	0.74	(0.69, 0.79)	0.01	(-0.03, 0.07)	-0.08	(-0.16, 0.01)
8	0.95	(0.90, 0.97)	0.72	(0.68, 0.77)	0.95	(0.87, 0.98)	0.79	(0.74, 0.82)	0.00	(-0.07, 0.07)	-0.07	(-0.13, 0.01)
9	0.91	(0.87, 0.94)	0.78	(0.74, 0.82)	0.93	(0.84, 0.97)	0.83	(0.80, 0.87)	-0.02	(-0.10, 0.08)	-0.05	(-0.11, 0.00)
10	0.87	(0.81, 0.92)	0.84	(0.80, 0.87)	0.91	(0.82, 0.96)	0.88	(0.85, 0.91)	-0.04	(-0.16, 0.07)	-0.04	(-0.10, 0.01)
11	0.85	--[d]	0.87	--[d]	0.84	(0.75, 0.90)	0.91	(0.89, 0.93)	0.01	(-0.15, 0.15)	-0.04	(-0.09, 0.01)
12	0.78	(0.72, 0.84)	0.90	(0.87, 0.92)	0.81	(0.71, 0.88)	0.94	(0.92, 0.95)	-0.03	(-0.19, 0.11)	-0.04	(-0.08, -0.00)
13	0.70	(0.65, 0.76)	0.92	(0.90, 0.94)	0.73	(0.62, 0.82)	0.95	(0.94, 0.97)	-0.03	(-0.24, 0.10)	-0.03	(-0.07, 0.00)
14	0.65	(0.58, 0.71)	0.94	(0.92, 0.96)	0.63	(0.51, 0.74)	0.97	(0.95, 0.98)	0.02	(-0.22, 0.20)	-0.03	(-0.06, -0.00)
15	0.58	(0.51, 0.65)	0.95	(0.93, 0.97)	0.54	(0.43, 0.65)	0.98	(0.96, 0.98)	0.04	(-0.21, 0.20)	-0.03	(-0.05, 0.00)

[a]N Studies = 26; N Participants = 4,132; N major depression = 629
[b]N Studies = 24; N Participants = 2,577; N major depression = 295
[c]10 bootstrap iterations (1%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E4. Comparison of PHQ-9 sensitivity and specificity estimates among women compared to men, among participants administered a semi-structured diagnostic interview

	Women[a]				Men[b]				Difference across groups[c] (Women – Men)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.99	(0.95, 1.00)	0.50	(0.43, 0.56)	0.98	(0.93, 1.00)	0.58	(0.53, 0.63)	0.01	(-0.03, 0.04)	-0.08	(-0.17, -0.01)
6	0.98	(0.95, 0.99)	0.59	(0.53, 0.65)	0.99	(0.92, 1.00)	0.66	(0.61, 0.70)	-0.01	(-0.04, 0.04)	-0.07	(-0.15, 0.01)
7	0.98	(0.94, 1.00)	0.66	(0.60, 0.72)	0.98	(0.91, 0.99)	0.72	(0.67, 0.76)	0.00	(-0.04, 0.07)	-0.06	(-0.13, 0.01)
8	0.97	(0.91, 0.99)	0.72	(0.67, 0.77)	0.94	(0.88, 0.97)	0.77	(0.74, 0.80)	0.03	(-0.06, 0.09)	-0.05	(-0.11, 0.01)
9	0.92	(0.86, 0.96)	0.78	(0.74, 0.82)	0.92	(0.86, 0.95)	0.83	(0.80, 0.85)	0.00	(-0.09, 0.10)	-0.05	(-0.10, 0.01)
10	0.91	(0.84, 0.95)	0.84	(0.79, 0.87)	0.86	(0.79, 0.90)	0.87	(0.85, 0.89)	0.05	(-0.07, 0.17)	-0.03	(-0.09, 0.01)
11	0.87	(0.80, 0.92)	0.87	(0.84, 0.90)	0.80	(0.73, 0.86)	0.90	(0.88, 0.92)	0.07	(-0.07, 0.21)	-0.03	(-0.08, 0.01)
12	0.81	(0.73, 0.87)	0.90	(0.87, 0.92)	0.75	(0.68, 0.82)	0.93	(0.91, 0.94)	0.06	(-0.11, 0.21)	-0.03	(-0.06, 0.01)
13	0.73	(0.66, 0.80)	0.92	(0.90, 0.94)	0.66	(0.59, 0.73)	0.94	(0.93, 0.96)	0.07	(-0.10, 0.23)	-0.02	(-0.06, 0.01)
14	0.68	(0.59, 0.76)	0.95	(0.92, 0.96)	0.60	(0.52, 0.67)	0.96	(0.94, 0.97)	0.08	(-0.09, 0.27)	-0.01	(-0.04, 0.01)
15	0.59	(0.50, 0.67)	0.96	(0.94, 0.97)	0.52	(0.44, 0.59)	0.97	(0.95, 0.98)	0.07	(-0.11, 0.25)	-0.01	(-0.04, 0.01)

[a]N Studies = 28; N Participants = 3,906; N major depression = 573
[b]N Studies = 25; N Participants = 2,812; N major depression = 351
[c]9 bootstrap iterations (0.9%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
Abbreviations: CI: confidence interval

Supplementary table E5. Comparison of PHQ-9 sensitivity and specificity estimates among participants from countries with a very high human development index compared to a high human development index, among participants administered a semi-structured diagnostic interview

	Very high human development index[a]				High human development index[b]				Difference across groups[c] (Very high human development index – high human development index)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.98	(0.95, 0.99)	0.56	(0.51, 0.61)	1.00	(0.68, 1.00)	0.45	(0.30, 0.62)	-0.02	(-0.04, 0.03)	0.11	(-0.05, 0.28)
6	0.97	(0.94, 0.99)	0.64	(0.59, 0.69)	1.00	(0.37, 1.00)	0.54	(0.36, 0.70)	-0.03	(-0.05, 0.04)	0.10	(-0.06, 0.30)
7	0.97	(0.92, 0.99)	0.71	(0.66, 0.75)	1.00	(0.23, 1.00)	0.62	(0.43, 0.78)	-0.03	(-0.07, 0.04)	0.09	(-0.07, 0.29)
8	0.94	(0.89, 0.97)	0.76	(0.73, 0.79)	0.99	(0.74, 1.00)	0.68	(0.48, 0.83)	-0.05	(-0.10, 0.05)	0.08	(-0.06, 0.28)
9	0.90	(0.85, 0.93)	0.81	(0.78, 0.84)	0.99	(0.75, 1.00)	0.76	(0.58, 0.88)	-0.09	(-0.15, 0.03)	0.05	(-0.08, 0.24)
10	0.86	(0.80, 0.90)	0.86	(0.83, 0.88)	0.99	(0.64, 1.00)	0.86	(0.65, 0.95)	-0.13	(-0.20, 0.00)	0.00	(-0.12, 0.19)
11	0.81	(0.75, 0.86)	0.89	(0.86, 0.91)	0.96	(0.80, 0.99)	0.89	(0.71, 0.96)	-0.15	(-0.24, 0.01)	0.00	(-0.09, 0.16)
12	0.76	(0.70, 0.81)	0.91	(0.89, 0.93)	0.88	(0.81, 0.92)	0.92	(0.77, 0.97)	-0.12	(-0.24, -0.01)	-0.01	(-0.08, 0.13)
13	0.68	(0.62, 0.74)	0.93	(0.92, 0.95)	0.77	--[d]	0.94	--[d]	-0.09	(-0.22, 0.05)	-0.01	(-0.07, 0.13)
14	0.63	(0.56, 0.69)	0.95	(0.94, 0.97)	0.74	(0.67, 0.80)	0.95	(0.79, 0.99)	-0.11	(-0.25, 0.04)	0.00	(-0.05, 0.13)
15	0.54	--[d]	0.96	--[d]	0.69	--[d]	0.96	--[d]	-0.15	(-0.31, -0.01)	0.00	(-0.04, 0.12)

[a]N Studies = 25; N Participants = 6,195; N major depression = 739
[b]N Studies = 4; N Participants = 530; N major depression = 185
[c]152 bootstrap iterations (15%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E6i. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and non-medical care settings, among participants administered a semi-structured diagnostic interview

	Primary care[a]				Non-medical care[b]				Difference across groups[c] (Primary care – non-medical care)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	1.00	(0.38, 1.00)	0.59	(0.48, 0.69)	0.95	(0.84, 0.99)	0.48	(0.40, 0.56)	0.05	(-0.01, 0.10)	0.11	(-0.04, 0.24)
6	1.00	(0.30, 1.00)	0.66	(0.56, 0.75)	0.95	(0.85, 0.98)	0.59	(0.52, 0.65)	0.05	(0.00, 0.11)	0.07	(-0.07, 0.19)
7	1.00	(0.64, 1.00)	0.73	(0.63, 0.81)	0.92	(0.82, 0.97)	0.66	(0.58, 0.73)	0.08	(0.01, 0.14)	0.07	(-0.06, 0.17)
8	0.99	(0.82, 1.00)	0.78	(0.69, 0.85)	0.89	(0.78, 0.95)	0.73	(0.66, 0.80)	0.10	(0.01, 0.17)	0.05	(-0.07, 0.14)
9	0.95	(0.90, 0.98)	0.83	(0.75, 0.89)	0.85	(0.77, 0.90)	0.82	(0.78, 0.85)	0.10	(0.02, 0.21)	0.01	(-0.08, 0.09)
10	0.94	(0.88, 0.97)	0.88	(0.79, 0.93)	0.82	(0.73, 0.88)	0.88	(0.85, 0.91)	0.12	(0.02, 0.23)	0.00	(-0.10, 0.07)
11	0.91	(0.82, 0.96)	0.91	(0.84, 0.95)	0.76	(0.67, 0.83)	0.92	(0.89, 0.94)	0.15	(0.00, 0.27)	-0.01	(-0.09, 0.04)
12	0.84	(0.78, 0.89)	0.92	(0.87, 0.96)	0.70	(0.60, 0.78)	0.94	(0.91, 0.96)	0.14	(-0.03, 0.26)	-0.02	(-0.08, 0.03)
13	0.77	(0.72, 0.82)	0.94	(0.89, 0.97)	0.62	(0.52, 0.71)	0.95	(0.93, 0.97)	0.15	(-0.11, 0.27)	-0.01	(-0.07, 0.03)
14	0.73	(0.66, 0.78)	0.96	(0.92, 0.98)	0.59	(0.49, 0.68)	0.97	(0.95, 0.98)	0.14	(-0.04, 0.27)	-0.01	(-0.06, 0.02)
15	0.65	(0.58, 0.72)	0.97	(0.93, 0.99)	0.43	(0.34, 0.52)	0.97	(0.95, 0.99)	0.22	(0.04, 0.37)	0.00	(-0.05, 0.02)

[a]N Studies = 9; N Participants = 3,163; N major depression = 377
[b]N Studies = 2; N Participants = 567; N major depression = 105
[c]212 bootstrap iterations (21.2%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
Abbreviations: CI: confidence interval

Supplementary table E6ii. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and inpatient speciality care settings, among participants administered a semi-structured diagnostic interview

	Primary care[a]				Inpatient specialty care[b]				Difference across groups[c] (Primary care – inpatient specialty care)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	1.00	(0.38, 1.00)	0.59	(0.48, 0.69)	1.00	(0.00, 1.00)	0.48	(0.36, 0.60)	0.00	(-0.03, 0.00)	0.11	(-0.08, 0.38)
6	1.00	(0.30, 1.00)	0.66	(0.56, 0.75)	1.00	(0.55, 1.00)	0.57	(0.45, 0.68)	0.00	(-0.03, 0.01)	0.09	(-0.08, 0.32)
7	1.00	(0.64, 1.00)	0.73	(0.63, 0.81)	1.00	(0.72, 1.00)	0.65	(0.58, 0.73)	0.00	(-0.03, 0.03)	0.08	(-0.08, 0.22)
8	0.99	(0.82, 1.00)	0.78	(0.69, 0.85)	0.96	(0.88, 0.99)	0.71	(0.64, 0.77)	0.03	(-0.06, 0.08)	0.07	(-0.06, 0.20)
9	0.95	(0.90, 0.98)	0.83	(0.75, 0.89)	0.95	(0.87, 0.98)	0.77	(0.73, 0.81)	0.00	(-0.08, 0.09)	0.06	(-0.05, 0.16)
10	0.94	(0.88, 0.97)	0.88	(0.79, 0.93)	0.92	(0.84, 0.96)	0.81	(0.78, 0.85)	0.02	(-0.10, 0.14)	0.07	(-0.04, 0.16)
11	0.91	(0.82, 0.96)	0.91	(0.84, 0.95)	0.90	(0.82, 0.95)	0.85	(0.81, 0.88)	0.01	(-0.14, 0.14)	0.06	(-0.04, 0.14)
12	0.84	(0.78, 0.89)	0.92	(0.87, 0.96)	0.86	(0.78, 0.92)	0.89	(0.85, 0.92)	-0.02	(-0.17, 0.15)	0.03	(-0.05, 0.11)
13	0.77	(0.72, 0.82)	0.94	(0.89, 0.97)	0.74	(0.65, 0.82)	0.91	(0.87, 0.94)	0.03	(-0.14, 0.25)	0.03	(-0.04, 0.10)
14	0.73	(0.66, 0.78)	0.96	(0.92, 0.98)	0.68	--[d]	0.93	--[d]	0.05	(-0.17, 0.38)	0.03	(-0.03, 0.09)
15	0.65	(0.58, 0.72)	0.97	(0.93, 0.99)	0.58	(0.35, 0.77)	0.94	(0.91, 0.97)	0.07	(-0.23, 0.60)	0.03	(-0.03, 0.07)

[a]N Studies = 9; N Participants = 3,163; N major depression = 377
[b]N Studies = 8; N Participants = 867; N major depression = 121
[c]407 bootstrap iterations (40.7%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E6iii. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and outpatient speciality care settings, among participants administered a semi-structured diagnostic interview

	Primary care[a]				Outpatient specialty care[b]				Difference across groups[c] (Primary care – outpatient specialty care)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	1.00	(0.38, 1.00)	0.59	(0.48, 0.69)	0.94	(0.89, 0.97)	0.53	(0.46, 0.60)	0.06	(-0.01, 0.09)	0.06	(-0.11, 0.21)
6	1.00	(0.30, 1.00)	0.66	(0.56, 0.75)	0.92	(0.86, 0.96)	0.61	(0.54, 0.68)	0.08	(-0.01, 0.12)	0.05	(-0.10, 0.19)
7	1.00	(0.64, 1.00)	0.73	(0.63, 0.81)	0.91	(0.83, 0.95)	0.68	(0.61, 0.74)	0.09	(-0.01, 0.15)	0.05	(-0.10, 0.17)
8	0.99	(0.82, 1.00)	0.78	(0.69, 0.85)	0.87	(0.79, 0.93)	0.74	(0.68, 0.79)	0.12	(-0.01, 0.20)	0.04	(-0.09, 0.14)
9	0.95	(0.90, 0.98)	0.83	(0.75, 0.89)	0.84	(0.75, 0.90)	0.79	(0.74, 0.83)	0.11	(-0.01, 0.22)	0.04	(-0.07, 0.13)
10	0.94	(0.88, 0.97)	0.88	(0.79, 0.93)	0.77	(0.67, 0.84)	0.84	(0.80, 0.88)	0.17	(0.00, 0.28)	0.04	(-0.08, 0.12)
11	0.91	(0.82, 0.96)	0.91	(0.84, 0.95)	0.72	(0.64, 0.79)	0.88	(0.84, 0.91)	0.19	(0.00, 0.33)	0.03	(-0.06, 0.10)
12	0.84	(0.78, 0.89)	0.92	(0.87, 0.96)	0.67	(0.58, 0.76)	0.90	(0.87, 0.93)	0.17	(-0.03, 0.31)	0.02	(-0.05, 0.08)
13	0.77	(0.72, 0.82)	0.94	(0.89, 0.97)	0.59	(0.49, 0.68)	0.93	(0.90, 0.95)	0.18	(0.02, 0.34)	0.01	(-0.06, 0.07)
14	0.73	(0.66, 0.78)	0.96	(0.92, 0.98)	0.54	(0.44, 0.64)	0.95	(0.92, 0.97)	0.19	(-0.02, 0.33)	0.01	(-0.05, 0.06)
15	0.65	(0.58, 0.72)	0.97	(0.93, 0.99)	0.49	(0.40, 0.58)	0.96	(0.93, 0.97)	0.16	(-0.04, 0.30)	0.01	(-0.03, 0.05)

[a]N Studies = 9; N Participants = 3,163; N major depression = 377
[b]N Studies = 12; N Participants = 2,128; N major depression = 321
[c]214 bootstrap iterations (21.4%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
Abbreviations: CI: confidence interval

Supplementary table E7. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 3 (Reference Standard) - Signalling Question 2 (*Were the reference standard results interpreted without knowledge of the results of the index test?*) , among participants administered a semi-structured diagnostic interview

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.98	(0.94, 0.99)	0.50	(0.43, 0.56)	0.98	(0.96, 0.99)	0.60	(0.53, 0.67)	0.00	(-0.04, 0.06)	-0.10	(-0.22, 0.01)
6	0.98	(0.93, 1.00)	0.58	(0.52, 0.64)	0.97	(0.93, 0.99)	0.68	(0.62, 0.74)	0.01	(-0.05, 0.07)	-0.10	(-0.21, 0.01)
7	0.98	(0.92, 1.00)	0.65	(0.59, 0.71)	0.96	(0.89, 0.99)	0.74	(0.69, 0.79)	0.02	(-0.06, 0.11)	-0.09	(-0.19, 0.00)
8	0.94	(0.90, 0.97)	0.71	(0.66, 0.76)	0.96	(0.85, 0.99)	0.79	(0.75, 0.83)	-0.02	(-0.09, 0.11)	-0.08	(-0.17, 0.00)
9	0.92	(0.87, 0.95)	0.77	(0.72, 0.81)	0.9	(0.83, 0.94)	0.84	(0.81, 0.87)	0.02	(-0.09, 0.14)	-0.07	(-0.15, 0.00)
10	0.90	(0.83, 0.94)	0.82	(0.77, 0.86)	0.86	(0.78, 0.91)	0.89	(0.86, 0.92)	0.04	(-0.11, 0.18)	-0.07	(-0.15, -0.01)
11	0.85	(0.78, 0.90)	0.85	(0.81, 0.89)	0.83	(0.73, 0.89)	0.92	(0.90, 0.94)	0.02	(-0.13, 0.20)	-0.07	(-0.14, -0.01)
12	0.80	(0.71, 0.86)	0.88	(0.85, 0.91)	0.77	(0.69, 0.83)	0.94	(0.92, 0.95)	0.03	(-0.12, 0.19)	-0.06	(-0.11, -0.01)
13	0.71	(0.63, 0.77)	0.91	(0.88, 0.94)	0.70	(0.63, 0.76)	0.95	(0.94, 0.97)	0.01	(-0.15, 0.16)	-0.04	(-0.10, 0.00)
14	0.65	(0.57, 0.73)	0.93	(0.90, 0.96)	0.65	(0.59, 0.70)	0.96	(0.96, 0.97)	0.00	(-0.15, 0.18)	-0.03	(-0.08, 0.00)
15	0.58	(0.49, 0.66)	0.95	(0.92, 0.97)	0.55	(0.45, 0.64)	0.97	(0.96, 0.98)	0.03	(-0.14, 0.28)	-0.02	(-0.07, 0.00)

[a]N Studies = 16; N Participants = 4,249; N major depression = 558
[b]N Studies = 13; N Participants = 2,476; N major depression = 366
[c]14 bootstrap iterations (1.4%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
Abbreviations: CI: confidence interval

Supplementary table E8. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 3 (Reference Standard) - Signalling Question 3 (*Did a qualified person administer the reference standard?*), among participants administered a semi-structured diagnostic interview

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.97	(0.92, 0.99)	0.55	(0.47, 0.62)	0.99	(0.96, 1.00)	0.54	(0.47, 0.61)	-0.02	(-0.08, 0.02)	0.01	(-0.12, 0.13)
6	0.96	(0.91, 0.98)	0.63	(0.56, 0.70)	0.99	(0.94, 1.00)	0.62	(0.55, 0.68)	-0.03	(-0.09, 0.02)	0.01	(-0.11, 0.13)
7	0.95	(0.88, 0.98)	0.69	(0.63, 0.76)	0.99	(0.90, 1.00)	0.69	(0.63, 0.75)	-0.04	(-0.12, 0.03)	0.00	(-0.11, 0.11)
8	0.93	(0.85, 0.97)	0.75	(0.69, 0.80)	0.96	(0.92, 0.98)	0.75	(0.70, 0.80)	-0.03	(-0.13, 0.06)	0.00	(-0.10, 0.09)
9	0.89	(0.81, 0.93)	0.80	(0.74, 0.84)	0.93	(0.88, 0.96)	0.81	(0.77, 0.84)	-0.04	(-0.15, 0.07)	-0.01	(-0.10, 0.06)
10	0.84	(0.76, 0.90)	0.85	(0.80, 0.89)	0.92	(0.85, 0.95)	0.86	(0.82, 0.89)	-0.08	(-0.20, 0.07)	-0.01	(-0.10, 0.06)
11	0.80	(0.73, 0.86)	0.88	(0.84, 0.92)	0.88	(0.79, 0.93)	0.89	(0.86, 0.92)	-0.08	(-0.22, 0.10)	-0.01	(-0.09, 0.05)
12	0.76	(0.68, 0.82)	0.90	(0.87, 0.93)	0.81	(0.73, 0.87)	0.92	(0.89, 0.94)	-0.05	(-0.21, 0.11)	-0.02	(-0.08, 0.04)
13	0.66	(0.58, 0.73)	0.93	(0.89, 0.95)	0.73	(0.67, 0.79)	0.94	(0.91, 0.95)	-0.07	(-0.24, 0.07)	-0.01	(-0.07, 0.03)
14	0.60	(0.51, 0.68)	0.95	(0.91, 0.97)	0.69	(0.61, 0.75)	0.95	(0.94, 0.97)	-0.09	(-0.26, 0.07)	0.00	(-0.06, 0.03)
15	0.54	--[d]	0.96	--[d]	0.58	(0.49, 0.67)	0.96	(0.95, 0.97)	-0.04	(-0.22, 0.18)	0.00	(-0.05, 0.02)

[a]N Studies = 14; N Participants = 3,462; N major depression = 433
[b]N Studies = 15; N Participants = 3,263; N major depression = 491
[c]30 bootstrap iterations (3%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E9. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 4 (Flow and Timing) - Signalling Question 4 (*Were all patients included in the analysis?)*, among participants administered a semi-structured diagnostic interview

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.97	(0.92, 0.99)	0.52	(0.45, 0.58)	0.99	(0.96, 1.00)	0.59	(0.51, 0.65)	-0.02	(-0.08, 0.01)	-0.07	(-0.19, 0.07)
6	0.96	(0.91, 0.99)	0.6	(0.53, 0.67)	0.99	(0.95, 1.00)	0.66	(0.60, 0.72)	-0.03	(-0.09, 0.02)	-0.06	(-0.17, 0.07)
7	0.96	(0.89, 0.99)	0.67	(0.61, 0.73)	0.99	(0.92, 1.00)	0.72	(0.66, 0.77)	-0.03	(-0.12, 0.04)	-0.05	(-0.16, 0.07)
8	0.94	(0.87, 0.98)	0.73	(0.67, 0.78)	0.96	(0.91, 0.98)	0.77	(0.73, 0.82)	-0.02	(-0.13, 0.07)	-0.04	(-0.14, 0.05)
9	0.9	(0.83, 0.95)	0.80	(0.75, 0.84)	0.93	(0.89, 0.96)	0.81	(0.77, 0.85)	-0.03	(-0.16, 0.07)	-0.01	(-0.10, 0.06)
10	0.88	(0.78, 0.93)	0.85	(0.80, 0.89)	0.90	(0.84, 0.94)	0.86	(0.82, 0.89)	-0.02	(-0.18, 0.10)	-0.01	(-0.09, 0.07)
11	0.84	(0.75, 0.90)	0.89	(0.85, 0.92)	0.85	(0.77, 0.91)	0.89	(0.85, 0.92)	-0.01	(-0.19, 0.14)	0.00	(-0.07, 0.07)
12	0.78	(0.70, 0.85)	0.91	(0.88, 0.94)	0.79	(0.72, 0.86)	0.91	(0.88, 0.93)	-0.01	(-0.19, 0.14)	0.00	(-0.06, 0.06)
13	0.70	(0.61, 0.77)	0.94	(0.90, 0.96)	0.71	(0.65, 0.77)	0.93	(0.91, 0.95)	-0.01	(-0.17, 0.15)	0.01	(-0.05, 0.06)
14	0.64	--[d]	0.95	--[d]	0.66	(0.59, 0.72)	0.95	(0.93, 0.96)	-0.02	(-0.20, 0.15)	0.00	(-0.04, 0.05)
15	0.54	--[d]	0.96	--[d]	0.59	(0.51, 0.66)	0.96	(0.94, 0.97)	-0.05	(-0.25, 0.15)	0.00	(-0.04, 0.04)

[a]N Studies = 17; N Participants = 2,579; N major depression = 499
[b]N Studies = 12; N Participants = 4,146; N major depression = 425
[c]49 bootstrap iterations (4.9%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CIs.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E10. Comparison of PHQ-9 sensitivity and specificity estimates among participants not currently diagnosed or receiving treatment for a mental health problem compared to all participants, among participants administered a fully structured diagnostic interview

	All participants[a]				Participants not currently diagnosed or receiving treatment for a mental health problem[b]				Difference across groups[c] (All participants – participants not currently diagnosed or receiving treatment for a mental health problem)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.93	(0.87, 0.97)	0.54	(0.43, 0.64)	0.95	(0.87, 0.98)	0.59	(0.42, 0.74)	-0.02	(-0.11, 0.05)	-0.05	(-0.20, 0.13)
6	0.91	(0.83, 0.95)	0.61	(0.51, 0.71)	0.94	(0.84, 0.98)	0.66	(0.48, 0.80)	-0.03	(-0.15, 0.04)	-0.05	(-0.18, 0.14)
7	0.86	(0.75, 0.92)	0.69	(0.59, 0.77)	0.91	(0.79, 0.97)	0.74	(0.60, 0.85)	-0.05	(-0.19, 0.05)	-0.05	(-0.17, 0.09)
8	0.82	(0.71, 0.89)	0.75	(0.66, 0.82)	0.88	(0.74, 0.95)	0.8	(0.67, 0.89)	-0.06	(-0.22, 0.06)	-0.05	(-0.15, 0.08)
9	0.74	(0.63, 0.83)	0.79	(0.72, 0.86)	0.79	(0.65, 0.89)	0.84	(0.71, 0.92)	-0.05	(-0.21, 0.09)	-0.05	(-0.13, 0.08)
10	0.70	(0.59, 0.80)	0.84	(0.77, 0.89)	0.76	(0.59, 0.87)	0.88	(0.76, 0.94)	-0.06	(-0.23, 0.11)	-0.04	(-0.11, 0.07)
11	0.62	(0.51, 0.72)	0.87	(0.81, 0.91)	0.65	(0.51, 0.77)	0.9	(0.80, 0.95)	-0.03	(-0.21, 0.15)	-0.03	(-0.09, 0.07)
12	0.57	(0.45, 0.68)	0.89	(0.85, 0.93)	0.60	(0.46, 0.73)	0.92	(0.84, 0.96)	-0.03	(-0.23, 0.14)	-0.03	(-0.07, 0.05)
13	0.49	(0.38, 0.61)	0.92	(0.89, 0.95)	0.55	(0.42, 0.67)	0.95	(0.89, 0.98)	-0.06	(-0.25, 0.12)	-0.03	(-0.07, 0.02)
14	0.44	(0.32, 0.56)	0.94	(0.91, 0.96)	0.48	(0.36, 0.61)	0.96	(0.92, 0.98)	-0.04	(-0.24, 0.14)	-0.02	(-0.06, 0.02)
15	0.35	(0.25, 0.46)	0.96	(0.93, 0.97)	0.42	(0.31, 0.53)	0.97	(0.94, 0.99)	-0.07	(-0.26, 0.09)	-0.01	(-0.04, 0.01)

[a]N Studies = 14; N Participants = 7,680; N major depression = 839
[b]N Studies = 6; N Participants = 4,161; N major depression = 306
[c]19 bootstrap iterations (2%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E11. Comparison of PHQ-9 sensitivity and specificity estimates among participants aged <60 compared to ≥60, among participants administered a fully structured diagnostic interview

	Age <60[a]				Age ≥60[b]				Difference across groups[c] (Age <60 – Age ≥60)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.94	(0.88, 0.97)	0.51	(0.41, 0.61)	0.94	(0.81, 0.98)	0.57	(0.43, 0.69)	0.00	(-0.08, 0.16)	-0.06	(-0.21, 0.12)
6	0.92	(0.84, 0.96)	0.59	(0.48, 0.69)	0.86	(0.74, 0.93)	0.63	(0.51, 0.74)	0.06	(-0.09, 0.23)	-0.04	(-0.18, 0.14)
7	0.87	(0.77, 0.93)	0.66	(0.57, 0.75)	0.78	(0.66, 0.87)	0.70	(0.60, 0.79)	0.09	(-0.13, 0.25)	-0.04	(-0.16, 0.12)
8	0.83	(0.72, 0.91)	0.73	(0.64, 0.80)	0.71	(0.60, 0.81)	0.78	(0.69, 0.85)	0.12	(-0.09, 0.32)	-0.05	(-0.16, 0.10)
9	0.76	(0.64, 0.85)	0.78	(0.69, 0.84)	0.64	(0.52, 0.75)	0.81	(0.73, 0.88)	0.12	(-0.12, 0.30)	-0.03	(-0.14, 0.10)
10	0.72	(0.60, 0.82)	0.82	(0.75, 0.88)	0.55	(0.44, 0.65)	0.86	(0.78, 0.91)	0.17	(-0.10, 0.37)	-0.04	(-0.13, 0.09)
11	0.64	(0.53, 0.74)	0.86	(0.80, 0.91)	0.46	(0.35, 0.56)	0.88	(0.81, 0.93)	0.18	(-0.12, 0.36)	-0.02	(-0.09, 0.08)
12	0.59	(0.47, 0.71)	0.88	(0.83, 0.92)	0.40	(0.31, 0.49)	0.91	(0.85, 0.95)	0.19	(-0.09, 0.38)	-0.03	(-0.09, 0.07)
13	0.52	(0.40, 0.64)	0.92	(0.87, 0.94)	0.31	(0.24, 0.40)	0.94	(0.89, 0.97)	0.21	(-0.08, 0.38)	-0.02	(-0.07, 0.05)
14	0.46	(0.34, 0.57)	0.94	(0.91, 0.96)	0.26	(0.19, 0.34)	0.95	(0.91, 0.97)	0.20	(-0.11, 0.41)	-0.01	(-0.05, 0.05)
15	0.38	(0.28, 0.49)	0.95	(0.93, 0.97)	0.20	(0.13, 0.30)	0.96	(0.93, 0.98)	0.18	(-0.10, 0.43)	-0.01	(-0.04, 0.04)

[a]N Studies = 14; N Participants = 5,504; N major depression = 645
[b]N Studies = 10; N Participants = 2,175; N major depression =194
[c]4 bootstrap iterations (0.4%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E12. Comparison of PHQ-9 sensitivity and specificity estimates among women compared to men, among participants administered a fully structured diagnostic interview

	Women[a]				Men[b]				Difference across groups[c] (Women – Men)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.92	(0.84, 0.97)	0.50	(0.39, 0.61)	0.93	(0.83, 0.97)	0.58	(0.48, 0.68)	-0.01	(-0.12, 0.10)	-0.08	(-0.19, -0.02)
6	0.89	(0.78, 0.95)	0.57	(0.44, 0.69)	0.92	(0.79, 0.97)	0.66	(0.56, 0.75)	-0.03	(-0.19, 0.09)	-0.09	(-0.22, -0.02)
7	0.83	(0.72, 0.91)	0.64	(0.51, 0.75)	0.85	(0.72, 0.92)	0.73	(0.65, 0.80)	-0.02	(-0.18, 0.13)	-0.09	(-0.28, -0.01)
8	0.79	(0.68, 0.87)	0.71	(0.59, 0.80)	0.82	(0.68, 0.91)	0.78	(0.71, 0.84)	-0.03	(-0.22, 0.15)	-0.07	(-0.21, -0.00)
9	0.72	(0.62, 0.80)	0.77	(0.66, 0.84)	0.73	(0.59, 0.83)	0.83	(0.76, 0.88)	-0.01	(-0.18, 0.16)	-0.06	(-0.14, -0.00)
10	0.67	(0.57, 0.76)	0.82	(0.73, 0.89)	0.72	(0.57, 0.83)	0.86	(0.80, 0.90)	-0.05	(-0.22, 0.13)	-0.04	(-0.12, 0.02)
11	0.60	(0.48, 0.70)	0.86	(0.78, 0.91)	0.62	(0.50, 0.73)	0.89	(0.84, 0.92)	-0.02	(-0.21, 0.14)	-0.03	(-0.09, 0.02)
12	0.55	(0.43, 0.66)	0.88	(0.82, 0.92)	0.57	(0.44, 0.68)	0.91	(0.87, 0.94)	-0.02	(-0.20, 0.16)	-0.03	(-0.08, 0.02)
13	0.48	(0.36, 0.59)	0.92	(0.87, 0.95)	0.49	(0.37, 0.61)	0.93	(0.90, 0.96)	-0.01	(-0.24, 0.17)	-0.01	(-0.07, 0.03)
14	0.43	(0.31, 0.55)	0.94	(0.90, 0.96)	0.42	(0.30, 0.55)	0.95	(0.92, 0.96)	0.01	(-0.21, 0.19)	-0.01	(-0.05, 0.02)
15	0.36	(0.26, 0.46)	0.95	(0.92, 0.97)	0.32	(0.21, 0.46)	0.97	(0.95, 0.98)	0.04	(-0.17, 0.22)	-0.02	(-0.05, 0.01)

[a]N Studies = 14; N Participants = 4,285; N major depression = 463
[b]N Studies = 13; N Participants = 3,395; N major depression =376
[c]5 bootstrap iterations (0.5%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E13i. Comparison of PHQ-9 sensitivity and specificity estimates among participants from countries with a very high human development index compared to a high human development index, among participants administered a fully structured diagnostic interview

	Very high human development index[a]				High human development index[b]				Difference across groups[c] (Very high human development index – high human development index)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.94	(0.90, 0.97)	0.49	(0.35, 0.64)	0.96	(0.28, 1.00)	0.58	(0.46, 0.70)	-0.02	(-0.08, 0.04)	-0.09	(-0.29, 0.08)
6	0.93	(0.87, 0.96)	0.56	(0.41, 0.70)	0.96	(0.17, 1.00)	0.70	(0.60, 0.79)	-0.03	(-0.11, 0.03)	-0.14	(-0.35, 0.01)
7	0.90	(0.81, 0.94)	0.64	(0.51, 0.76)	0.96	(0.16, 1.00)	0.77	(0.67, 0.84)	-0.06	(-0.17, 0.02)	-0.13	(-0.31, 0.01)
8	0.86	(0.76, 0.92)	0.71	(0.58, 0.81)	0.96	(0.10, 1.00)	0.84	(0.73, 0.91)	-0.10	(-0.24, -0.00)	-0.13	(-0.31, -0.02)
9	0.80	(0.69, 0.88)	0.75	(0.63, 0.84)	0.72	(0.39, 0.91)	0.89	(0.82, 0.94)	0.08	(-0.11, 0.24)	-0.14	(-0.31, -0.04)
10	0.78	(0.65, 0.87)	0.80	(0.70, 0.88)	0.63	(0.38, 0.83)	0.92	(0.84, 0.96)	0.15	(-0.07, 0.32)	-0.12	(-0.27, -0.03)
11	0.69	(0.56, 0.79)	0.84	(0.76, 0.90)	0.54	(0.30, 0.77)	0.94	(0.88, 0.97)	0.15	(-0.08, 0.32)	-0.10	(-0.22, -0.03)
12	0.65	(0.51, 0.76)	0.87	(0.80, 0.92)	0.51	(0.31, 0.70)	0.95	(0.91, 0.98)	0.14	(-0.09, 0.33)	-0.08	(-0.18, -0.03)
13	0.57	(0.43, 0.69)	0.90	(0.85, 0.94)	0.45	(0.23, 0.69)	0.99	(0.84, 1.00)	0.12	(-0.09, 0.33)	-0.09	(-0.16, -0.04)
14	0.51	(0.37, 0.65)	0.92	(0.88, 0.95)	0.40	(0.18, 0.67)	0.99	(0.87, 1.00)	0.11	(-0.09, 0.37)	-0.07	(-0.13, -0.04)
15	0.43	(0.31, 0.55)	0.94	(0.91, 0.96)	0.29	(0.13, 0.54)	0.99	(0.93, 1.00)	0.14	(-0.06, 0.35)	-0.05	(-0.10, -0.03)

[a]N Studies = 9; N Participants = 5,740; N major depression = 592
[b]N Studies = 2; N Participants = 326; N major depression = 61
[c]738 bootstrap iterations (74%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E13ii. Comparison of PHQ-9 sensitivity and specificity estimates among participants from countries with a very high human development index compared to a low-medium human development index, among participants administered a fully structured diagnostic interview

	Very high human development index[a]				Low-medium human development index[b]				Difference across groups[c] (Very high human development index – low-medium human development index)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.94	(0.90, 0.97)	0.49	(0.35, 0.64)	0.79	(0.58, 0.91)	0.63	(0.46, 0.77)	0.15	(-0.04, 0.33)	-0.14	(-0.38, 0.06)
6	0.93	(0.87, 0.96)	0.56	(0.41, 0.70)	0.70	(0.50, 0.84)	0.71	(0.55, 0.83)	0.23	(-0.02, 0.46)	-0.15	(-0.39, 0.05)
7	0.90	(0.81, 0.94)	0.64	(0.51, 0.76)	0.59	(0.38, 0.76)	0.76	(0.61, 0.86)	0.31	(0.11, 0.56)	-0.12	(-0.33, 0.06)
8	0.86	(0.76, 0.92)	0.71	(0.58, 0.81)	0.56	(0.39, 0.72)	0.80	(0.68, 0.89)	0.30	(0.09, 0.53)	-0.09	(-0.31, 0.05)
9	0.80	(0.69, 0.88)	0.75	(0.63, 0.84)	0.50	(0.32, 0.68)	0.84	(0.73, 0.91)	0.30	(0.05, 0.55)	-0.09	(-0.29, 0.04)
10	0.78	(0.65, 0.87)	0.80	(0.70, 0.88)	0.47	(0.32, 0.62)	0.88	(0.77, 0.94)	0.31	(0.03, 0.57)	-0.08	(-0.27, 0.04)
11	0.69	(0.56, 0.79)	0.84	(0.76, 0.90)	0.43	(0.30, 0.57)	0.90	(0.81, 0.95)	0.26	(0.02, 0.52)	-0.06	(-0.20, 0.03)
12	0.65	(0.51, 0.76)	0.87	(0.80, 0.92)	0.35	(0.22, 0.51)	0.92	(0.84, 0.96)	0.30	(0.06, 0.65)	-0.05	(-0.17, 0.03)
13	0.57	(0.43, 0.69)	0.90	(0.85, 0.94)	0.29	(0.17, 0.44)	0.93	(0.88, 0.97)	0.28	(0.01, 0.58)	-0.03	(-0.12, 0.02)
14	0.51	(0.37, 0.65)	0.92	(0.88, 0.95)	0.24	(0.14, 0.37)	0.95	(0.92, 0.97)	0.27	(0.04, 0.54)	-0.03	(-0.09, 0.01)
15	0.43	(0.31, 0.55)	0.94	(0.91, 0.96)	0.16	(0.05, 0.42)	0.97	(0.94, 0.98)	0.27	(0.05, 0.50)	-0.03	(-0.08, 0.01)

[a]N Studies = 9; N Participants = 5,740; N major depression = 592
[b]N Studies = 3; N Participants = 1,614; N major depression = 186
[c]738 bootstrap iterations (74%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E14i. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and non-medical care settings, among participants administered a fully structured diagnostic interview

	Primary care[a]				Non-medical care[b]				Difference across groups[c] (Primary care – non-medical care)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.94	(0.80, 0.99)	0.58	(0.49, 0.66)	0.90	(0.69, 0.97)	0.69	(0.65, 0.71)	0.04	(-0.10, 0.12)	-0.11	(-0.20, -0.00)
6	0.91	(0.77, 0.97)	0.68	(0.61, 0.75)	0.87	(0.69, 0.95)	0.72	(0.65, 0.79)	0.04	(-0.14, 0.15)	-0.04	(-0.15, 0.04)
7	0.85	(0.70, 0.93)	0.74	(0.67, 0.80)	0.79	(0.65, 0.88)	0.78	(0.70, 0.84)	0.06	(-0.18, 0.22)	-0.04	(-0.13, 0.06)
8	0.84	(0.63, 0.94)	0.81	(0.73, 0.86)	0.75	(0.55, 0.88)	0.82	(0.74, 0.88)	0.09	(-0.17, 0.25)	-0.01	(-0.09, 0.06)
9	0.75	(0.63, 0.84)	0.85	(0.79, 0.90)	0.65	(0.48, 0.78)	0.85	(0.76, 0.91)	0.10	(-0.07, 0.28)	0.00	(-0.07, 0.07)
10	0.71	(0.60, 0.80)	0.88	(0.84, 0.92)	0.61	(0.44, 0.75)	0.88	(0.80, 0.93)	0.10	(-0.07, 0.31)	0.00	(-0.06, 0.06)
11	0.65	(0.52, 0.76)	0.91	(0.87, 0.94)	0.51	(0.35, 0.67)	0.91	(0.83, 0.95)	0.14	(-0.07, 0.29)	0.00	(-0.05, 0.04)
12	0.60	(0.52, 0.68)	0.93	(0.89, 0.95)	0.44	(0.28, 0.62)	0.92	(0.84, 0.96)	0.16	(-0.03, 0.32)	0.01	(-0.04, 0.05)
13	0.53	(0.44, 0.63)	0.95	(0.90, 0.98)	0.37	(0.19, 0.59)	0.94	(0.89, 0.97)	0.16	(-0.04, 0.36)	0.01	(-0.04, 0.06)
14	0.47	(0.37, 0.57)	0.96	(0.93, 0.98)	0.33	(0.17, 0.53)	0.95	(0.91, 0.98)	0.14	(-0.06, 0.34)	0.01	(-0.03, 0.05)
15	0.39	(0.29, 0.50)	0.97	(0.94, 0.99)	0.26	(0.13, 0.44)	0.96	(0.93, 0.98)	0.13	(-0.11, 0.29)	0.01	(-0.03, 0.03)

[a]N Studies = 5; N Participants = 3,578; N major depression = 273
[b]N Studies = 2; N Participants = 963; N major depression = 74
[c]901 bootstrap iterations (90%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E14ii. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and inpatient speciality care settings, among participants administered a fully structured diagnostic interview

	Primary care[a]				Inpatient specialty care[b]				Difference across groups[c] (Primary care – inpatient specialty care)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.94	(0.80, 0.99)	0.58	(0.49, 0.66)	0.99	(0.40, 1.00)	0.33	(0.18, 0.51)	-0.05	(-0.19, 0.02)	0.25	(0.16, 0.35)
6	0.91	(0.77, 0.97)	0.68	(0.61, 0.75)	0.99	(0.44, 1.00)	0.37	(0.24, 0.54)	-0.08	(-0.23, 0.01)	0.31	(0.23, 0.39)
7	0.85	(0.70, 0.93)	0.74	(0.67, 0.80)	0.94	(0.79, 0.99)	0.47	(0.28, 0.66)	-0.09	(-0.29, 0.05)	0.27	(0.19, 0.37)
8	0.84	(0.63, 0.94)	0.81	(0.73, 0.86)	0.92	(0.74, 0.98)	0.56	(0.38, 0.72)	-0.08	(0.29, 0.10)	0.25	(0.17, 0.33)
9	0.75	(0.63, 0.84)	0.85	(0.79, 0.90)	0.89	(0.68, 0.97)	0.61	(0.45, 0.75)	-0.14	(-0.29, 0.03)	0.24	(0.17, 0.31)
10	0.71	(0.60, 0.80)	0.88	(0.84, 0.92)	0.89	(0.68, 0.97)	0.69	(0.54, 0.80)	-0.18	(-0.03, -0.02)	0.19	(0.14, 0.26)
11	0.65	(0.52, 0.76)	0.91	(0.87, 0.94)	0.83	(0.48, 0.97)	0.73	(0.60, 0.83)	-0.18	(-0.36, 0.03)	0.18	(0.12, 0.23)
12	0.60	(0.52, 0.68)	0.93	(0.89, 0.95)	0.83	(0.48, 0.96)	0.77	(0.68, 0.85)	-0.23	(-0.41, -0.07)	0.16	(0.09, 0.20)
13	0.53	(0.44, 0.63)	0.95	(0.90, 0.98)	0.71	(0.33, 0.93)	0.83	(0.70, 0.92)	-0.18	(-0.39, 0.05)	0.12	(0.05, 0.17)
14	0.47	(0.37, 0.57)	0.96	(0.93, 0.98)	0.69	(0.27, 0.93)	0.86	(0.75, 0.93)	-0.22	(-0.48, -0.00)	0.10	(0.05, 0.15)
15	0.39	(0.29, 0.50)	0.97	(0.94, 0.99)	0.6	(0.31, 0.83)	0.90	(0.81, 0.95)	-0.21	(-0.43, 0.04)	0.07	(0.03, 0.11)

[a]N Studies = 5; N Participants = 3,578; N major depression = 273
[b]N Studies = 2; N Participants = 372; N major depression = 34
[c]901 bootstrap iterations (90%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E14iii. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and outpatient speciality care settings, among participants administered a fully structured diagnostic interview

	Primary care[a]				Outpatient specialty care[b]				Difference across groups[c] (Primary care – outpatient specialty care)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.94	(0.80, 0.99)	0.58	(0.49, 0.66)	0.91	(0.76, 0.97)	0.52	(0.29, 0.74)	0.03	(-0.15, 0.27)	0.06	(-0.14, 0.29)
6	0.91	(0.77, 0.97)	0.68	(0.61, 0.75)	0.87	(0.66, 0.96)	0.59	(0.35, 0.79)	0.04	(-0.18, 0.33)	0.09	(-0.08, 0.31)
7	0.85	(0.70, 0.93)	0.74	(0.67, 0.80)	0.83	(0.54, 0.96)	0.67	(0.46, 0.83)	0.02	(-0.27, 0.40)	0.07	(-0.07, 0.24)
8	0.84	(0.63, 0.94)	0.81	(0.73, 0.86)	0.77	(0.50, 0.92)	0.72	(0.52, 0.86)	0.07	(-0.21, 0.42)	0.09	(-0.04, 0.26)
9	0.75	(0.63, 0.84)	0.85	(0.79, 0.90)	0.69	(0.46, 0.86)	0.76	(0.57, 0.89)	0.06	(-0.21, 0.40)	0.09	(-0.03, 0.24)
10	0.71	(0.60, 0.80)	0.88	(0.84, 0.92)	0.63	(0.38, 0.83)	0.80	(0.62, 0.91)	0.08	(-0.20, 0.38)	0.08	(-0.02, 0.22)
11	0.65	(0.52, 0.76)	0.91	(0.87, 0.94)	0.54	(0.34, 0.73)	0.85	(0.70, 0.93)	0.11	(-0.16, 0.35)	0.06	(-0.02, 0.17)
12	0.60	(0.52, 0.68)	0.93	(0.89, 0.95)	0.50	(0.28, 0.71)	0.88	(0.75, 0.94)	0.10	(-0.19, 0.43)	0.05	(-0.02, 0.15)
13	0.53	(0.44, 0.63)	0.95	(0.90, 0.98)	0.42	(0.22, 0.65)	0.91	(0.83, 0.95)	0.11	(-0.20, 0.41)	0.04	(-0.01, 0.12)
14	0.47	(0.37, 0.57)	0.96	(0.93, 0.98)	0.36	(0.18, 0.59)	0.93	(0.87, 0.96)	0.11	(-0.22, 0.36)	0.03	(-0.01, 0.09)
15	0.39	(0.29, 0.50)	0.97	(0.94, 0.99)	0.30	(0.14, 0.52)	0.95	(0.90, 0.98)	0.09	(-0.16, 0.41)	0.02	(-0.02, 0.06)

[a]N Studies = 5; N Participants = 3,578; N major depression = 273
[b]N Studies = 5; N Participants = 2,767; N major depression = 458
[c]901 bootstrap iterations (90%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E15. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 1 (Participant Selection) - Signalling Question 1 (*Was a consecutive or random sample of participants enrolled?*), among participants administered a fully structured diagnostic interview

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.95	(0.70, 0.99)	0.68	(0.58, 0.76)	0.93	(0.86, 0.96)	0.47	(0.35, 0.59)	0.02	(-0.16, 0.12)	0.21	(0.05, 0.39)
6	0.92	(0.67, 0.98)	0.76	(0.68, 0.82)	0.91	(0.82, 0.96)	0.55	(0.42, 0.66)	0.01	(-0.25, 0.15)	0.21	(0.07, 0.39)
7	0.83	(0.46, 0.97)	0.81	(0.75, 0.86)	0.86	(0.76, 0.92)	0.63	(0.51, 0.73)	-0.03	(-0.41, 0.19)	0.18	(0.06, 0.34)
8	0.82	(0.43, 0.97)	0.86	(0.82, 0.89)	0.82	(0.70, 0.90)	0.69	(0.59, 0.78)	0.00	(-0.39, 0.25)	0.17	(0.06, 0.31)
9	0.70	(0.47, 0.86)	0.89	(0.85, 0.92)	0.75	(0.63, 0.84)	0.74	(0.64, 0.82)	-0.05	(-0.39, 0.16)	0.15	(0.05, 0.28)
10	0.69	(0.51, 0.83)	0.92	(0.89, 0.94)	0.72	(0.58, 0.83)	0.79	(0.70, 0.86)	-0.03	(-0.38, 0.17)	0.13	(0.05, 0.25)
11	0.63	(0.49, 0.76)	0.93	(0.91, 0.95)	0.63	(0.49, 0.75)	0.83	(0.76, 0.89)	0.00	(-0.35, 0.20)	0.10	(0.04, 0.20)
12	0.55	(0.38, 0.70)	0.95	(0.93, 0.96)	0.59	(0.45, 0.72)	0.86	(0.80, 0.91)	-0.04	(-0.42, 0.17)	0.09	(0.03, 0.16)
13	0.48	(0.30, 0.67)	0.96	(0.93, 0.98)	0.50	(0.37, 0.64)	0.90	(0.85, 0.93)	-0.02	(-0.43, 0.21)	0.06	(0.02, 0.13)
14	0.48	(0.40, 0.55)	0.97	(0.95, 0.99)	0.45	(0.31, 0.59)	0.92	(0.89, 0.95)	0.03	(-0.40, 0.22)	0.05	(0.01, 0.10)
15	0.32	(0.14, 0.58)	0.98	(0.97, 0.98)	0.37	(0.26, 0.49)	0.94	(0.91, 0.96)	-0.05	(-0.49, 0.18)	0.04	(0.01, 0.08)

[a]N Studies = 4; N Participants = 3,360; N major depression = 211
[b]N Studies = 10; N Participants = 4,320; N major depression = 628
[c]102 bootstrap iterations (10%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E16. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 3 (Reference Standard) - Signalling Question 2 (*Were the reference standard results interpreted without knowledge of the results of the index test?*), among participants administered a fully structured diagnostic interview

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.93	(0.81, 0.98)	0.62	(0.52, 0.70)	0.93	--[d]	0.42	--[d]	0.00	(-0.14, 0.12)	0.20	(-0.03, 0.41)
6	0.90	--[d]	0.70	--[d]	0.92	(0.80, 0.97)	0.49	(0.31, 0.67)	-0.02	(-0.20, 0.14)	0.21	(-0.00, 0.43)
7	0.82	(0.67, 0.91)	0.76	(0.69, 0.82)	0.89	(0.74, 0.96)	0.57	(0.41, 0.72)	-0.07	(-0.28, 0.13)	0.19	(0.00, 0.39)
8	0.78	(0.62, 0.89)	0.81	(0.75, 0.86)	0.86	(0.68, 0.94)	0.64	(0.48, 0.78)	-0.08	(-0.28, 0.18)	0.17	(0.01, 0.36)
9	0.71	(0.57, 0.81)	0.85	(0.80, 0.89)	0.78	(0.61, 0.89)	0.69	(0.53, 0.82)	-0.07	(-0.31, 0.14)	0.16	(0.01, 0.35)
10	0.67	(0.54, 0.78)	0.89	(0.85, 0.92)	0.75	(0.55, 0.88)	0.74	(0.59, 0.85)	-0.08	(-0.32, 0.16)	0.15	(0.02, 0.32)
11	0.59	(0.46, 0.70)	0.91	(0.87, 0.94)	0.67	(0.47, 0.82)	0.80	(0.67, 0.88)	-0.08	(-0.35, 0.18)	0.11	(0.01, 0.25)
12	0.53	(0.42, 0.64)	0.93	(0.89, 0.95)	0.64	(0.42, 0.81)	0.83	(0.73, 0.90)	-0.11	(-0.41, 0.16)	0.10	(0.01, 0.20)
13	0.46	(0.36, 0.57)	0.95	(0.92, 0.97)	0.56	(0.34, 0.75)	0.87	(0.80, 0.92)	-0.10	(-0.41, 0.18)	0.08	(0.01, 0.17)
14	0.40	(0.30, 0.51)	0.96	(0.94, 0.97)	0.51	(0.29, 0.72)	0.91	(0.85, 0.95)	-0.11	(-0.42, 0.17)	0.05	(0.00, 0.13)
15	0.33	(0.24, 0.44)	0.97	(0.95, 0.98)	0.40	(0.23, 0.59)	0.93	(0.89, 0.96)	-0.07	(-0.39, 0.16)	0.04	(-0.00, 0.10)

[a]N Studies = 8; N Participants = 5,140; N major depression = 522
[b]N Studies = 6; N Participants = 2,540; N major depression = 317
[c]19 bootstrap iterations (2%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E17. Comparison of PHQ-9 sensitivity and specificity estimates among participants not currently diagnosed or receiving treatment for a mental health problem compared to all participants, among participants administered the MINI

	All participants[a]				Participants not currently diagnosed or receiving treatment for a mental health problem[b]				Difference across groups[c] (All participants – participants not currently diagnosed or receiving treatment for a mental health problem)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.96	(0.93, 0.98)	0.57	(0.50, 0.64)	0.94	(0.86, 0.98)	0.63	(0.54, 0.70)	0.02	(-0.03, 0.12)	-0.06	(-0.18, 0.06)
6	0.93	(0.87, 0.97)	0.66	(0.59, 0.72)	0.92	(0.82, 0.96)	0.72	(0.64, 0.78)	0.01	(-0.06, 0.15)	-0.06	(-0.17, 0.04)
7	0.90	(0.82, 0.94)	0.72	(0.66, 0.78)	0.89	(0.73, 0.96)	0.78	(0.72, 0.83)	0.01	(-0.09, 0.20)	-0.06	(-0.16, 0.03)
8	0.86	(0.78, 0.91)	0.78	(0.73, 0.83)	0.84	(0.68, 0.93)	0.83	(0.78, 0.87)	0.02	(-0.09, 0.23)	-0.05	(-0.13, 0.03)
9	0.82	(0.72, 0.88)	0.84	(0.79, 0.87)	0.77	(0.58, 0.89)	0.89	(0.85, 0.92)	0.05	(-0.11, 0.27)	-0.05	(-0.12, 0.00)
10	0.77	(0.68, 0.83)	0.87	(0.83, 0.90)	0.71	(0.59, 0.81)	0.91	(0.88, 0.94)	0.06	(-0.09, 0.24)	-0.04	(-0.11, 0.01)
11	0.70	(0.62, 0.77)	0.90	(0.86, 0.92)	0.62	(0.55, 0.70)	0.94	(0.92, 0.95)	0.08	(-0.08, 0.23)	-0.04	(-0.10, -0.00)
12	0.65	(0.56, 0.72)	0.92	(0.89, 0.94)	0.59	(0.47, 0.69)	0.96	(0.94, 0.97)	0.06	(-0.11, 0.24)	-0.04	(-0.08, -0.00)
13	0.57	(0.49, 0.65)	0.94	(0.91, 0.96)	0.48	(0.39, 0.58)	0.97	(0.95, 0.98)	0.09	(-0.11, 0.23)	-0.03	(-0.07, 0.00)
14[d]	0.49	(0.42, 0.56)	0.96	(0.93, 0.97)	0.4	(0.31, 0.50)	0.97	(0.96, 0.98)	0.09	(-0.11, 0.22)	-0.01	(-0.05, 0.01)
15[d]	0.42	(0.35, 0.49)	0.97	(0.95, 0.98)	0.34	(0.25, 0.46)	0.98	(0.97, 0.99)	0.08	(-0.12, 0.22)	-0.01	(-0.04, 0.01)

[a]N Studies = 15; N Participants = 2,952; N major depression = 549
[b]N Studies = 6; N Participants = 927; N major depression = 168
[c]4 bootstrap iterations (0.4%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
[d]For these cutoffs, among all participants, the default optimizer in glmer failed, thus bobyqa was used instead.
Abbreviations: CI: confidence interval

Supplementary table E18. Comparison of PHQ-9 sensitivity and specificity estimates among participants aged <60 compared to ≥60, among participants administered the MINI

	Age <60[a]				Age ≥60[b]				Difference across groups[c] (Age <60 – Age ≥60)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.97	(0.93, 0.98)	0.52	(0.45, 0.59)	0.97	(0.88, 0.99)	0.65	(0.58, 0.72)	0.00	(-0.06, 0.12)	-0.13	(-0.27, 0.04)
6	0.95	(0.92, 0.98)	0.61	(0.54, 0.67)	0.88	(0.76, 0.95)	0.72	(0.66, 0.78)	0.07	(-0.05, 0.24)	-0.11	(-0.23, 0.03)
7	0.93	(0.86, 0.96)	0.68	(0.62, 0.74)	0.85	(0.73, 0.93)	0.79	(0.73, 0.83)	0.08	(-0.07, 0.24)	-0.11	(-0.21, 0.02)
8	0.88	(0.81, 0.93)	0.75	(0.69, 0.80)	0.83	(0.71, 0.91)	0.84	(0.79, 0.88)	0.05	(-0.12, 0.21)	-0.09	(-0.21, 0.02)
9	0.84	(0.74, 0.90)	0.81	(0.76, 0.85)	0.80	(0.67, 0.88)	0.87	(0.83, 0.91)	0.04	(-0.16, 0.24)	-0.06	(-0.15, 0.02)
10	0.79	(0.70, 0.85)	0.85	(0.80, 0.88)	0.75	(0.64, 0.84)	0.90	(0.86, 0.94)	0.04	(-0.17, 0.18)	-0.05	(-0.14, 0.02)
11	0.70	(0.61, 0.77)	0.88	(0.84, 0.91)	0.71	(0.59, 0.81)	0.92	(0.89, 0.95)	-0.01	(-0.24, 0.15)	-0.04	(-0.12, 0.02)
12	0.65	(0.55, 0.74)	0.91	(0.87, 0.93)	0.62	(0.52, 0.70)	0.94	(0.90, 0.96)	0.03	(-0.19, 0.22)	-0.03	(-0.10, 0.03)
13	0.58	(0.49, 0.67)	0.93	(0.90, 0.95)	0.52	(0.43, 0.60)	0.97	(0.92, 0.98)	0.06	(-0.21, 0.23)	-0.04	(-0.09, 0.02)
14	0.51	(0.44, 0.59)	0.95	(0.93, 0.97)	0.42	(0.35, 0.50)	0.97	(0.93, 0.99)	0.09	(-0.15, 0.23)	-0.02	(-0.06, 0.03)
15	0.43	(0.35, 0.51)	0.96	(0.94, 0.98)	0.37	(0.30, 0.44)	0.98	(0.95, 0.99)	0.06	(-0.11, 0.22)	-0.02	(-0.05, 0.01)

[a]N Studies = 14; N Participants = 1,958; N major depression =310
[b]N Studies = 13; N Participants =979; N major depression =239
[c]8 bootstrap iterations (0.8%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E19. Comparison of PHQ-9 sensitivity and specificity estimates among women compared to men, among participants administered the MINI

	Women[a]				Men[b]				Difference across groups[c] (Women – Men)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.96	(0.92, 0.98)	0.47	(0.37, 0.57)	0.99	(0.91, 1.00)	0.63	(0.54, 0.72)	-0.03	(-0.08, 0.03)	-0.16	(-0.43, -0.03)
6	0.93	(0.84, 0.97)	0.56	(0.45, 0.66)	0.95	(0.89, 0.98)	0.72	(0.63, 0.79)	-0.02	(-0.14, 0.06)	-0.16	(-0.42, -0.01)
7	0.90	(0.80, 0.96)	0.64	(0.54, 0.72)	0.92	(0.84, 0.96)	0.78	(0.71, 0.84)	-0.02	(-0.14, 0.11)	-0.14	(-0.32, -0.03)
8	0.87	(0.77, 0.93)	0.71	(0.63, 0.78)	0.87	(0.77, 0.93)	0.84	(0.78, 0.89)	0.00	(-0.17, 0.15)	-0.13	(-0.28, -0.04)
9	0.81	(0.71, 0.89)	0.78	(0.72, 0.83)	0.83	(0.71, 0.90)	0.87	(0.82, 0.91)	-0.02	(-0.19, 0.15)	-0.09	(-0.21, -0.01)
10	0.77	(0.68, 0.84)	0.82	(0.76, 0.87)	0.77	(0.66, 0.85)	0.90	(0.85, 0.94)	0.00	(-0.16, 0.20)	-0.08	(-0.17, -0.00)
11	0.68	(0.59, 0.76)	0.86	(0.81, 0.90)	0.73	--[d]	0.92	--[d]	-0.05	(-0.21, 0.17)	-0.06	(-0.14, 0.00)
12	0.64	(0.54, 0.72)	0.9	(0.85, 0.93)	0.65	(0.53, 0.75)	0.93	(0.90, 0.96)	-0.01	(-0.21, 0.21)	-0.03	(-0.10, 0.01)
13	0.57	--[d]	0.93	--[d]	0.55	(0.44, 0.65)	0.95	(0.92, 0.97)	0.02	(-0.17, 0.23)	-0.02	(-0.08, 0.02)
14	0.48	(0.40, 0.57)	0.95	(0.91, 0.97)	0.47	(0.38, 0.56)	0.96	(0.93, 0.97)	0.01	(-0.20, 0.23)	-0.01	(-0.06, 0.02)
15	0.41	(0.34, 0.48)	0.96	(0.93, 0.98)	0.40	(0.30, 0.50)	0.98	(0.95, 0.99)	0.01	(-0.16, 0.20)	-0.02	(-0.05, 0.01)

[a]N Studies = 15; N Participants = 1,666; N major depression = 337
[b]N Studies = 15; N Participants = 1,286; N major depression = 212
[c]20 bootstrap iterations (0.2%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E20i. Comparison of PHQ-9 sensitivity and specificity estimates among participants from countries with a very high human development index compared to a high human development index, among participants administered the MINI

	Very high human development index[a]				High human development index[b]				Difference across groups[c] (Very high human development index – high human development index)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.97	(0.93, 0.99)	0.61	(0.51, 0.70)	0.94	(0.75, 0.99)	0.50	(0.40, 0.61)	0.03	(-0.04, 0.17)	0.11	(-0.12, 0.24)
6	0.93	(0.83, 0.97)	0.69	(0.60, 0.77)	0.89	(0.77, 0.95)	0.59	(0.48, 0.69)	0.04	(-0.08, 0.17)	0.10	(-0.10, 0.24)
7	0.90	(0.79, 0.95)	0.75	(0.67, 0.82)	0.85	(0.69, 0.94)	0.65	(0.55, 0.74)	0.05	(-0.10, 0.23)	0.10	(-0.07, 0.22)
8	0.86	(0.76, 0.93)	0.81	(0.74, 0.86)	0.78	(0.62, 0.89)	0.72	(0.64, 0.79)	0.08	(-0.07, 0.30)	0.09	(-0.06, 0.18)
9	0.82	(0.69, 0.90)	0.85	(0.79, 0.90)	0.73	(0.56, 0.85)	0.80	(0.75, 0.84)	0.09	(-0.09, 0.34)	0.05	(-0.07, 0.12)
10	0.77	(0.65, 0.86)	0.88	(0.82, 0.92)	0.69	(0.56, 0.79)	0.85	(0.81, 0.88)	0.08	(-0.08, 0.30)	0.03	(-0.07, 0.10)
11	0.70	(0.58, 0.79)	0.90	(0.85, 0.94)	0.67	(0.55, 0.78)	0.89	(0.85, 0.91)	0.03	(-0.16, 0.26)	0.01	(-0.07, 0.08)
12	0.65	(0.53, 0.75)	0.92	(0.88, 0.95)	0.67	(0.55, 0.78)	0.90	(0.87, 0.93)	-0.02	(-0.22, 0.22)	0.02	(-0.05, 0.08)
13	0.57	--[d]	0.94	--[d]	0.59	(0.46, 0.71)	0.94	(0.91, 0.95)	-0.02	(-0.20, 0.21)	0.00	(-0.07, 0.06)
14	0.49	--[d]	0.96	--[d]	0.49	(0.37, 0.62)	0.95	(0.93, 0.97)	0.00	(-0.16, 0.22)	0.01	(-0.05, 0.06)
15	0.43	(0.34, 0.52)	0.97	(0.94, 0.99)	0.43	(0.31, 0.55)	0.97	(0.95, 0.98)	0.00	(-0.17, 0.24)	0.00	(-0.04, 0.03)

[a]N Studies = 10; N Participants = 1,924; N major depression = 430
[b]N Studies = 3; N Participants = 542; N major depression = 61
[c]708 bootstrap iterations (71%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E20ii. Comparison of PHQ-9 sensitivity and specificity estimates among participants from countries with a very high human development index compared to a low-medium human development index, among participants administered the MINI

	Very high human development index[a]				Low-medium human development index[b]				Difference across groups[c] (Very high human development index – low-medium human development index)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.97	(0.93, 0.99)	0.61	(0.51, 0.70)	0.97	(0.87, 0.99)	0.49	(0.44, 0.53)	0.00	(-0.05, 0.06)	0.12	(-0.06, 0.25)
6	0.93	(0.83, 0.97)	0.69	(0.60, 0.77)	0.97	(0.87, 0.99)	0.58	(0.53, 0.63)	-0.04	(-0.13, 0.05)	0.11	(-0.04, 0.21)
7	0.90	(0.79, 0.95)	0.75	(0.67, 0.82)	0.93	(0.83, 0.97)	0.67	(0.62, 0.71)	-0.03	(-0.16, 0.07)	0.08	(-0.06, 0.17)
8	0.86	(0.76, 0.93)	0.81	(0.74, 0.86)	0.90	(0.79, 0.95)	0.73	(0.69, 0.77)	-0.04	(-0.16, 0.09)	0.08	(-0.05, 0.15)
9	0.82	(0.69, 0.90)	0.85	(0.79, 0.90)	0.88	(0.77, 0.94)	0.80	(0.76, 0.84)	-0.06	(-0.23, 0.08)	0.05	(-0.08, 0.10)
10	0.77	(0.65, 0.86)	0.88	(0.82, 0.92)	0.83	(0.71, 0.90)	0.84	(0.81, 0.87)	-0.06	(-0.21, 0.11)	0.04	(-0.10, 0.09)
11	0.70	(0.58, 0.79)	0.9	(0.85, 0.94)	0.71	(0.58, 0.81)	0.87	(0.83, 0.90)	-0.01	(-0.18, 0.19)	0.03	(-0.09, 0.09)
12	0.65	(0.53, 0.75)	0.92	(0.88, 0.95)	0.59	(0.46, 0.70)	0.90	(0.86, 0.92)	0.06	(-0.16, 0.27)	0.02	(-0.06, 0.07)
13	0.57	--[d]	0.94	--[d]	0.52	(0.39, 0.64)	0.93	(0.91, 0.95)	0.05	(-0.19, 0.26)	0.01	(-0.09, 0.05)
14	0.49	--[d]	0.96	--[d]	0.45	(0.25, 0.67)	0.96	(0.91, 0.98)	0.04	(-0.16, 0.26)	0.00	(-0.07, 0.04)
15	0.43	(0.34, 0.52)	0.97	(0.94, 0.99)	0.34	(0.17, 0.56)	0.97	(0.94, 0.98)	0.09	(-0.14, 0.29)	0.00	(-0.05, 0.03)

[a]N Studies = 10; N Participants = 1,924; N major depression = 430
[b]N Studies = 2; N Participants = 486; N major depression = 58
[c]708 bootstrap iterations (71%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E21i. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and non-medical care settings, among participants administered the MINI

	Primary care[a]				Non-medical care[b]				Difference across groups[c] (Primary care – non-medical care)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.98	(0.93, 0.99)	0.54	(0.43, 0.64)	0.95	(0.77, 0.99)	0.42	(0.22, 0.65)	0.03	(-0.04, 0.10)	0.12	(-0.09, 0.27)
6	0.91	(0.73, 0.98)	0.63	(0.52, 0.73)	0.95	(0.78, 0.99)	0.54	(0.35, 0.72)	-0.04	(-0.20, 0.07)	0.09	(-0.10, 0.21)
7	0.89	(0.69, 0.96)	0.69	(0.59, 0.77)	0.90	(0.69, 0.98)	0.59	(0.40, 0.76)	-0.01	(-0.22, 0.12)	0.10	(-0.08, 0.20)
8	0.83	(0.64, 0.93)	0.76	(0.68, 0.82)	0.87	(0.66, 0.96)	0.68	(0.51, 0.81)	-0.04	(-0.29, 0.14)	0.08	(-0.08, 0.16)
9	0.81	(0.63, 0.91)	0.82	(0.77, 0.85)	0.85	(0.67, 0.94)	0.74	(0.56, 0.87)	-0.04	(-0.29, 0.14)	0.08	(-0.05, 0.15)
10	0.74	(0.56, 0.86)	0.86	(0.82, 0.89)	0.84	(0.68, 0.93)	0.77	(0.60, 0.88)	-0.10	(-0.31, 0.11)	0.09	(-0.02, 0.16)
11	0.67	(0.48, 0.82)	0.88	(0.84, 0.91)	0.82	(0.68, 0.91)	0.80	(0.60, 0.92)	-0.15	(-0.37, 0.09)	0.08	(-0.02, 0.15)
12	0.61	(0.42, 0.78)	0.90	(0.87, 0.93)	0.82	(0.68, 0.91)	0.85	(0.68, 0.93)	-0.21	(-0.46, 0.05)	0.05	(-0.03, 0.12)
13	0.54	(0.38, 0.68)	0.94	(0.91, 0.95)	0.75	(0.56, 0.88)	0.87	(0.66, 0.95)	-0.21	(-0.42, 0.05)	0.07	(-0.01, 0.12)
14	0.47	(0.35, 0.59)	0.96	(0.94, 0.97)	0.63	(0.45, 0.78)	0.89	(0.73, 0.96)	-0.16	(-0.38, 0.09)	0.07	(0.01, 0.11)
15	0.38	(0.27, 0.50)	0.97	(0.96, 0.98)	0.57	(0.37, 0.75)	0.92	(0.79, 0.98)	-0.19	(-0.38, 0.04)	0.05	(-0.00, 0.08)

[a]N Studies = 5; N Participants = 1,290; N major depression = 168
[b]N Studies = 2; N Participants = 299; N major depression = 72
[c]589 bootstrap iterations (59%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E21ii. Comparison of PHQ-9 sensitivity and specificity estimates among participants from primary care and inpatient or outpatient speciality care settings, among participants administered the MINI

	Primary care[a]				Inpatient or outpatient specialty care[b]				Difference across groups[c] (Primary care – inpatient or outpatient specialty care)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.98	(0.93, 0.99)	0.54	(0.43, 0.64)	0.96	(0.90, 0.98)	0.63	(0.53, 0.71)	0.02	(-0.05, 0.10)	-0.09	(-0.32, 0.08)
6	0.91	(0.73, 0.98)	0.63	(0.52, 0.73)	0.94	(0.85, 0.97)	0.70	(0.62, 0.77)	-0.03	(-0.19, 0.14)	-0.07	(-0.28, 0.05)
7	0.89	(0.69, 0.96)	0.69	(0.59, 0.77)	0.90	(0.79, 0.96)	0.77	(0.70, 0.83)	-0.01	(-0.24, 0.17)	-0.08	(-0.27, 0.03)
8	0.83	(0.64, 0.93)	0.76	(0.68, 0.82)	0.87	(0.75, 0.93)	0.82	(0.76, 0.87)	-0.04	(-0.24, 0.18)	-0.06	(-0.23, 0.03)
9	0.81	(0.63, 0.91)	0.82	(0.77, 0.85)	0.81	(0.65, 0.90)	0.87	(0.82, 0.91)	0.00	(-0.23, 0.26)	-0.05	(-0.17, 0.02)
10	0.74	(0.56, 0.86)	0.86	(0.82, 0.89)	0.75	(0.63, 0.84)	0.90	(0.85, 0.93)	-0.01	(-0.25, 0.25)	-0.04	(-0.15, 0.03)
11	0.67	(0.48, 0.82)	0.88	(0.84, 0.91)	0.67	(0.58, 0.74)	0.92	(0.88, 0.95)	0.00	(-0.22, 0.29)	-0.04	(-0.13, 0.02)
12	0.61	(0.42, 0.78)	0.90	(0.87, 0.93)	0.61	(0.54, 0.67)	0.94	(0.90, 0.96)	0.00	(-0.27, 0.30)	-0.04	(-0.11, 0.02)
13	0.54	(0.38, 0.68)	0.94	(0.91, 0.95)	0.53	(0.46, 0.60)	0.96	(0.92, 0.98)	0.01	(-0.25, 0.25)	-0.02	(-0.08, 0.03)
14	0.47	(0.35, 0.59)	0.96	(0.94, 0.97)	0.46	(0.39, 0.54)	0.97	(0.94, 0.98)	0.01	(-0.25, 0.21)	-0.01	(-0.06, 0.02)
15	0.38	(0.27, 0.50)	0.97	(0.96, 0.98)	0.39	(0.32, 0.47)	0.98	(0.95, 0.99)	-0.01	(-0.25, 0.19)	-0.01	(-0.04, 0.02)

[a]N Studies = 5; N Participants = 1,290; N major depression = 168
[b]N Studies = 8; N Participants = 1,363; N major depression = 309
[c]589 bootstrap iterations (59%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E22. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 1 (Participant Selection) - Signalling Question 1 (*Was a consecutive or random sample of participants enrolled?*), among participants administered the MINI

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.92	(0.85, 0.96)	0.64	(0.53, 0.74)	0.98	(0.94, 0.99)	0.53	(0.44, 0.62)	-0.06	(-0.15, 0.03)	0.11	(-0.04, 0.29)
6	0.89	(0.78, 0.95)	0.72	(0.63, 0.80)	0.94	(0.87, 0.98)	0.62	(0.53, 0.69)	-0.05	(-0.20, 0.07)	0.10	(-0.02, 0.27)
7	0.85	(0.75, 0.91)	0.79	(0.71, 0.85)	0.92	(0.82, 0.96)	0.68	(0.61, 0.75)	-0.07	(-0.24, 0.08)	0.11	(-0.01, 0.24)
8	0.83	(0.72, 0.90)	0.84	(0.78, 0.89)	0.88	(0.77, 0.94)	0.74	(0.68, 0.80)	-0.05	(-0.24, 0.12)	0.10	(0.01, 0.21)
9	0.76	(0.63, 0.86)	0.88	(0.83, 0.91)	0.84	(0.72, 0.92)	0.81	(0.75, 0.85)	-0.08	(-0.28, 0.12)	0.07	(-0.00, 0.17)
10	0.73	(0.62, 0.81)	0.91	(0.87, 0.94)	0.79	(0.68, 0.87)	0.84	(0.79, 0.88)	-0.06	(-0.26, 0.13)	0.07	(0.00, 0.16)
11	0.66	(0.55, 0.76)	0.93	(0.90, 0.96)	0.72	(0.61, 0.80)	0.87	(0.82, 0.91)	-0.06	(-0.28, 0.12)	0.06	(0.01, 0.15)
12	0.62	(0.49, 0.74)	0.95	(0.92, 0.96)	0.66	(0.56, 0.75)	0.90	(0.85, 0.93)	-0.04	(-0.28, 0.17)	0.05	(0.00, 0.12)
13	0.55	(0.41, 0.69)	0.97	(0.94, 0.98)	0.59	(0.49, 0.68)	0.92	(0.88, 0.95)	-0.04	(-0.27, 0.18)	0.05	(0.00, 0.11)
14	0.47	(0.35, 0.60)	0.98	(0.95, 0.99)	0.50	(0.41, 0.58)	0.94	(0.91, 0.96)	-0.03	(-0.23, 0.19)	0.04	(0.00, 0.09)
15	0.40	(0.28, 0.52)	0.98	(0.97, 0.99)	0.43	(0.34, 0.52)	0.96	(0.93, 0.97)	-0.03	(-0.23, 0.17)	0.02	(-0.00, 0.07)

[a]N Studies = 5; N Participants = 1,085; N major depression = 155
[b]N Studies = 10; N Participants = 1,867; N major depression = 394
[c]55 bootstrap iterations (6%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E23. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 3 (Reference Standard) - Signalling Question 2 (*Were the reference standard results interpreted without knowledge of the results of the index test?*), among participants administered the MINI

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.98	(0.93, 0.99)	0.60	(0.51, 0.68)	0.93	(0.84, 0.97)	0.49	(0.37, 0.62)	0.05	(-0.03, 0.14)	0.11	(-0.06, 0.28)
6	0.94	(0.85, 0.98)	0.68	(0.60, 0.75)	0.93	(0.82, 0.97)	0.58	(0.47, 0.68)	0.01	(-0.10, 0.14)	0.10	(-0.04, 0.25)
7	0.90	(0.80, 0.96)	0.75	(0.68, 0.81)	0.89	(0.77, 0.95)	0.64	(0.54, 0.73)	0.01	(-0.13, 0.18)	0.11	(-0.02, 0.24)
8	0.87	(0.77, 0.93)	0.81	(0.75, 0.85)	0.85	(0.70, 0.93)	0.70	(0.62, 0.78)	0.02	(-0.13, 0.22)	0.11	(-0.01, 0.21)
9	0.82	(0.70, 0.90)	0.86	(0.82, 0.89)	0.82	(0.64, 0.92)	0.76	(0.66, 0.84)	0.00	(-0.17, 0.24)	0.10	(0.00, 0.20)
10	0.75	(0.65, 0.83)	0.89	(0.86, 0.92)	0.81	(0.65, 0.91)	0.78	(0.70, 0.85)	-0.06	(-0.23, 0.19)	0.11	(0.03, 0.21)
11	0.67	(0.58, 0.76)	0.91	(0.89, 0.94)	0.75	(0.62, 0.85)	0.82	(0.72, 0.89)	-0.08	(-0.26, 0.15)	0.09	(0.01, 0.20)
12	0.62	(0.53, 0.70)	0.93	(0.91, 0.95)	0.71	(0.56, 0.83)	0.85	(0.77, 0.91)	-0.09	(-0.30, 0.15)	0.08	(0.01, 0.17)
13	0.55	(0.46, 0.63)	0.95	(0.93, 0.96)	0.64	(0.48, 0.77)	0.88	(0.78, 0.93)	-0.09	(-0.30, 0.16)	0.07	(0.00, 0.17)
14	0.47	(0.39, 0.55)	0.97	(0.96, 0.97)	0.55	(0.42, 0.67)	0.89	(0.82, 0.93)	-0.08	(-0.27, 0.14)	0.08	(0.02, 0.15)
15	0.39	(0.32, 0.46)	0.98	(0.97, 0.98)	0.49	(0.36, 0.63)	0.92	(0.85, 0.96)	-0.10	(-0.29, 0.10)	0.06	(-0.00, 0.13)

[a]N Studies = 11; N Participants = 2,413; N major depression = 427
[b]N Studies = 4; N Participants = 539; N major depression = 122
[c]82 bootstrap iterations (8%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table E24. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "unclear" risk of bias for QUADAS-2 Domain 4 (Flow and Timing) - Signalling Question 1 (*Was there an appropriate interval between index test and reference standard?)*, among participants administered the MINI

	Low risk of bias[a]				Unclear risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear risk of bias)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.97	(0.93, 0.98)	0.53	(0.43, 0.63)	0.97	(0.83, 1.00)	0.63	(0.56, 0.70)	0.00	(-0.05, 0.11)	-0.10	(-0.26, 0.15)
6	0.95	(0.90, 0.98)	0.62	(0.52, 0.71)	0.85	--[d]	0.69	--[d]	0.10	(-0.07, 0.28)	-0.07	(-0.23, 0.13)
7	0.93	(0.86, 0.96)	0.69	(0.59, 0.77)	0.82	(0.62, 0.93)	0.75	(0.71, 0.79)	0.11	(-0.11, 0.31)	-0.06	(-0.22, 0.10)
8	0.89	(0.81, 0.94)	0.75	(0.66, 0.83)	0.77	(0.59, 0.88)	0.80	(0.76, 0.83)	0.12	(-0.12, 0.37)	-0.05	(-0.20, 0.09)
9	0.86	(0.86, 0.86)	0.81	(0.81, 0.81)	0.71	(0.57, 0.81)	0.86	(0.82, 0.89)	0.15	(-0.16, 0.35)	-0.05	(-0.20, 0.06)
10	0.80	(0.70, 0.87)	0.85	(0.76, 0.90)	0.69	(0.55, 0.80)	0.89	(0.83, 0.92)	0.11	(-0.22, 0.28)	-0.04	(-0.19, 0.07)
11	0.72	(0.63, 0.80)	0.88	(0.81, 0.92)	0.64	(0.53, 0.74)	0.93	(0.88, 0.96)	0.08	(-0.25, 0.21)	-0.05	(-0.17, 0.04)
12	0.67	(0.57, 0.76)	0.90	(0.84, 0.94)	0.59	(0.46, 0.71)	0.94	(0.91, 0.97)	0.08	(-0.30, 0.29)	-0.04	(-0.13, 0.04)
13	0.61	(0.51, 0.70)	0.92	(0.87, 0.96)	0.48	(0.36, 0.60)	0.97	(0.92, 0.99)	0.13	(-0.38, 0.38)	-0.05	(-0.13, 0.02)
14	0.52	(0.43, 0.60)	0.95	(0.90, 0.97)	0.39	(0.31, 0.47)	0.97	(0.93, 0.99)	0.13	(-0.47, 0.45)	-0.02	(-0.10, 0.03)
15	0.44	(0.36, 0.52)	0.96	(0.93, 0.98)	0.33	--[d]	0.98	--[d]	0.11	(-0.56, 0.36)	-0.02	(-0.06, 0.02)

[a]N Studies = 13; N Participants = 2,346; N major depression = 394
[b]N Studies = 5; N Participants = 606; N major depression = 155
[c]41 bootstrap iterations (4%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E25. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 4 (Flow and Timing) - Signalling Question 2 (*Did all patients receive a reference standard?)*, among participants administered the MINI

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.97	(0.93, 0.99)	0.57	(0.49, 0.64)	0.94	(0.86, 0.98)	0.59	(0.40, 0.76)	0.03	(-0.05, 0.13)	-0.02	(-0.23, 0.16)
6	0.94	(0.86, 0.98)	0.65	(0.59, 0.72)	0.91	(0.77, 0.97)	0.67	(0.49, 0.82)	0.03	(-0.06, 0.18)	-0.02	(-0.20, 0.15)
7	0.91	(0.81, 0.96)	0.72	(0.65, 0.77)	0.88	(0.75, 0.95)	0.75	(0.57, 0.87)	0.03	(-0.11, 0.17)	-0.03	(-0.19, 0.13)
8	0.87	(0.76, 0.93)	0.78	(0.72, 0.82)	0.85	(0.74, 0.92)	0.81	(0.65, 0.91)	0.02	(-0.13, 0.19)	-0.03	(-0.16, 0.11)
9	0.84	(0.72, 0.91)	0.82	(0.78, 0.86)	0.77	(0.61, 0.87)	0.87	(0.76, 0.93)	0.07	(-0.11, 0.26)	-0.05	(-0.14, 0.06)
10	0.79	(0.68, 0.87)	0.86	(0.81, 0.89)	0.72	(0.60, 0.82)	0.90	(0.82, 0.95)	0.07	(-0.11, 0.24)	-0.04	(-0.13, 0.03)
11	0.72	(0.61, 0.80)	0.88	(0.84, 0.92)	0.64	--[d]	0.93	--[d]	0.08	(-0.09, 0.29)	-0.05	(-0.12, 0.03)
12	0.68	(0.57, 0.77)	0.91	(0.87, 0.94)	0.56	(0.47, 0.64)	0.94	(0.88, 0.97)	0.12	(-0.07, 0.31)	-0.03	(-0.11, 0.03)
13	0.61	(0.51, 0.70)	0.93	(0.89, 0.95)	0.47	(0.38, 0.56)	0.97	(0.91, 0.99)	0.14	(-0.07, 0.33)	-0.04	(-0.10, 0.01)
14	0.53	(0.45, 0.61)	0.95	(0.92, 0.97)	0.37	(0.30, 0.45)	0.97	(0.93, 0.99)	0.16	(-0.02, 0.33)	-0.02	(-0.07, 0.01)
15	0.47	--[d]	0.96	--[d]	0.28	(0.22, 0.36)	0.98	(0.95, 0.99)	0.19	(0.03, 0.36)	-0.02	(-0.06, 0.01)

[a]N Studies = 11; N Participants = 1,962; N major depression = 393
[b]N Studies = 4; N Participants = 990; N major depression = 156
[c]115 bootstrap iterations (12%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
[d]Model for this cutoff did not converge.
Abbreviations: CI: confidence interval

Supplementary table E26. Comparison of PHQ-9 sensitivity and specificity estimates among studies and participants categorized as having "low" risk of bias compared to "high" or "unclear" risk of bias for QUADAS-2 Domain 4 (Flow and Timing) - Signalling Question 4 (*Were all patients included in the analysis?)*, among participants administered the MINI

	Low risk of bias[a]				Unclear or high risk of bias[b]				Difference across groups[c] (Low risk of bias – unclear or high risk of bias)			
Cutoff	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**	**Sensitivity**	**95% CI**	**Specificity**	**95% CI**
5	0.97	(0.93, 0.99)	0.54	(0.45, 0.63)	0.95	(0.90, 0.98)	0.66	(0.56, 0.75)	0.02	(-0.06, 0.09)	-0.12	(-0.27, 0.23)
6	0.95	(0.90, 0.98)	0.64	(0.55, 0.71)	0.85	(0.66, 0.94)	0.72	(0.61, 0.80)	0.10	(-0.06, 0.29)	-0.08	(-0.21, 0.17)
7	0.92	(0.86, 0.96)	0.71	(0.62, 0.78)	0.81	(0.60, 0.92)	0.77	(0.69, 0.83)	0.11	(-0.09, 0.33)	-0.06	(-0.17, 0.17)
8	0.89	(0.81, 0.93)	0.78	(0.71, 0.83)	0.78	(0.59, 0.90)	0.80	(0.72, 0.86)	0.11	(-0.11, 0.32)	-0.02	(-0.12, 0.17)
9	0.85	(0.76, 0.91)	0.83	(0.78, 0.87)	0.72	(0.52, 0.85)	0.85	(0.76, 0.91)	0.13	(-0.12, 0.34)	-0.02	(-0.11, 0.15)
10	0.79	(0.71, 0.86)	0.87	(0.82, 0.91)	0.70	(0.50, 0.84)	0.87	(0.79, 0.92)	0.09	(-0.15, 0.30)	0.00	(-0.08, 0.16)
11	0.73	(0.65, 0.81)	0.90	(0.85, 0.93)	0.61	(0.50, 0.70)	0.90	(0.82, 0.94)	0.12	(-0.14, 0.29)	0.00	(-0.08, 0.13)
12	0.69	(0.59, 0.78)	0.92	(0.88, 0.94)	0.54	(0.47, 0.61)	0.92	(0.85, 0.96)	0.15	(-0.14, 0.32)	0.00	(-0.07, 0.11)
13	0.62	(0.51, 0.71)	0.94	(0.91, 0.96)	0.46	(0.39, 0.53)	0.94	(0.86, 0.98)	0.16	(-0.09, 0.32)	0.00	(-0.06, 0.10)
14	0.53	(0.44, 0.62)	0.96	(0.93, 0.97)	0.39	(0.32, 0.47)	0.95	(0.88, 0.98)	0.14	(-0.08, 0.29)	0.01	(-0.04, 0.10)
15	0.46	(0.37, 0.55)	0.97	(0.95, 0.98)	0.33	(0.26, 0.40)	0.96	(0.89, 0.99)	0.13	(-0.08, 0.28)	0.01	(-0.03, 0.09)

[a]N Studies = 11; N Participants = 2,270; N major depression = 353
[b]N Studies = 4; N Participants = 682; N major depression = 196
[c]121 bootstrap iterations (12%) did not produce a difference estimate for all cutoffs (5-15). These iterations were removed prior to determining the bootstrapped CI.
Abbreviations: CI: confidence interval

Supplementary table F. QUADAS-2 ratings for each primary study included in the present study

First Author, Year	Domain 1: Participant Selection					Domain 2: Index Test				Domain 3: Reference Standard					Domain 4: Flow and Timing				
	SQ1	SQ2	SQ3	RoB	AC	SQ 1	SQ2	RoB	AC	SQ1	SQ2	SQ3	RoB	AC	SQ1	SQ2	SQ3	SQ4	RoB
Semi-structured Interviews																			
Amoozegar, 2017[1a]	U/C	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	U/C	U/C	Low	U/C	Yes	Yes	No	U/C
Ayalon, 2010[2]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	U/C	U/C	Low	Yes	Yes	Yes	Yes	Low
Beraldi, 2014[3]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	U/C	U/C	Low	Yes	Yes	Yes	Yes	Low
Bombardier, 2012[4]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	IPD[b]	Yes	Yes	U/C	IPD[b]
Chagas, 2013[5]	Yes	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	No	U/C
Eack, 2006[6]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Fann, 2005[7]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	No	Yes	High	Low	Yes	U/C	Yes	No	High
Fiest, 2014[8]	U/C	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	U/C	U/C	Low	U/C	Yes	Yes	No	U/C
Fischer, 2014[9]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Gjerdingen, 2009[10]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	U/C	U/C	Low	U/C	Yes	Yes	U/C	U/C
Gräfe, 2004[11]	Yes	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	U/C	U/C	Low	Yes	Yes	Yes	U/C	U/C
Khamseh, 2011[12]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Kwan, 2012[13]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	U/C	U/C	U/C	Yes	Yes	Yes	U/C	U/C
Lambert, 2015[14a]	No	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Liu, 2011[15]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	No	U/C
McGuire, 2013[16]	U/C	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	U/C	U/C	Low	Yes	Yes	Yes	Yes	Low
Osório, 2009[17]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	U/C	U/C	Low	Yes	Yes	Yes	Yes	Low
Osório, 2012[18]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	U/C	U/C	Low	Yes	Yes	Yes	Yes	Low
Picardi, 2005[19]	Yes	Yes	Yes	Low	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Richardson, 2010[20]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	U/C	U/C	Low	Yes	Yes	Yes	Yes	Low
Rooney, 2013[21]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Sidebottom, 2012[22]	No	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	No	High	Low	IPD[b]	Yes	Yes	No	U/C
Simning, 2012[23]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	No	High	Low	Yes	Yes	Yes	Yes	Low
Turner, Unpublished	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Turner, 2012[24]	U/C	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Twist, 2013[25]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	No	Yes	High	Low	Yes	Yes	Yes	U/C	U/C
Vöhringer, 2013[26]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	U/C	U/C	Low	Yes	Yes	Yes	Yes	Low
Williams, 2012[27]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	IPD[b]	Yes	Yes	Yes	IPD[b]
Wittkampf, 2009[28]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	U/C	U/C	Low	Yes	Yes	Yes	No	U/C

First Author, Year	Domain 1: Participant Selection					Domain 2: Index Test				Domain 3: Reference Standard					Domain 4: Flow and Timing				
	SQ1	SQ2	SQ3	RoB	AC	SQ 1	SQ2	RoB	AC	SQ1	SQ2	SQ3	RoB	AC	SQ1	SQ2	SQ3	SQ4	RoB
Fully Structured Interviews																			
Arroll, 2010[29]	Yes	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Azah, 2005[30]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	U/C	Yes	U/C	Yes	U/C	U/C
de Man-van Ginkel, 2012[31]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Delgadillo, 2011[32]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Gelaye, 2014[33]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Hahn, 2006[34]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	U/C	Yes	Yes	Yes	U/C
Henkel, 2004[35]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Hobfoll, 2011[36]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	U/C	Yes	Yes	Yes	U/C
Kiely, 2014[37]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	U/C	U/C	Yes	U/C	U/C
Mohd Sidik, 2012[38]	Yes	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	U/C	Yes	Yes	Yes	Yes	Low
Patel, 2008[39]	Yes	Yes	Yes	Low	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Pence, 2012[40]	Yes	Yes	Yes	Low	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Razykov, 2013[41]	No	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Thombs, 2008[42]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Mini International Neuropsychiatric Interviews (MINI)																			
Akena, 2013[43]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Cholera, 2014[44]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	U/C	Yes	Yes	Yes	Low	U/C	Yes	No	Yes	Yes	Low
Hides, 2007[45]	No	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	Yes	Yes	Yes	Yes	Low
Hyphantis, 2011[46]	Yes	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	U/C	U/C	Yes	U/C	U/C
Hyphantis, 2014[47]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Inagaki, 2013[48]	Yes	Yes	Yes	Low	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	No	Yes	Yes	High
Lamers, 2008[49]	U/C	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	IPD[b]	Yes	Yes	No	U/C
Lotrakul, 2008[50]	No	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	No	Yes	Yes	High
Muramatsu, 2007[51]	U/C	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Persoons, 2001[52]	Yes	Yes	Yes	Low	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
Santos, 2013[53]	Yes	Yes	Yes	Low	Low	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	U/C	Yes	Yes	Yes	U/C
Stafford, 2007[54]	No	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	U/C	Low
Sung, 2013[55]	Yes	Yes	Yes	Low	U/C	N/A	N/A	Low	Low	Yes	Yes	Yes	Low	Low	Yes	Yes	Yes	Yes	Low
van Steenbergen-Weijenburg, 2010[56]	No	Yes	Yes	U/C	U/C	N/A	N/A	Low	Low	Yes	No	Yes	High	Low	IPD[b]	Yes	Yes	No	High
Zhang, 2013[57]	U/C	Yes	Yes	U/C	Low	N/A	N/A	Low	Low	Yes	U/C	Yes	U/C	Low	IPD[b]	Yes	Yes	Yes	IPD[b]

Abbreviations: AC: acceptability concern, RoB: risk of bias, SQ: signalling question, N/A: not applicable; U/C: Unclear

[a]Was unpublished at the time of electronic database search

[b]Rating varies at the individual participant level

SUPPLEMENTARY MATERIAL REFERENCES

1. 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.
2. Ayalon L, Goldfracht M, Bech P. 'Do you think you suffer from depression?' Re-evaluating the use of a single item question for the screening of depression in older primary care patients. *Int J Geriatr Psychiatry*. 2010;**25**:497–502.
3. Beraldi A, Baklayan A, Hoster E, Hiddemann W, Heussner P. Which questionnaire is most suitable for the detection of depressive disorders in haemato-oncological patients? Comparison between HADS, CES-D and PHQ-9. *Oncol Res Treat*. 2014;**37**:108–109.
4. Bombardier CH, Kalpakjian CZ, Graves DE, Dyer JR, Tate DG, Fann JR. Validity of the Patient Health Questionnaire-9 in assessing major depressive disorder during inpatient spinal cord injury rehabilitation. *Arch Phys Med Rehabil*. 2012;**93**:1838–1845.
5. Chagas MH, Tumas V, Rodrigues GR, et al. Validation and internal consistency of Patient Health Questionnaire-9 for major depression in Parkinson's disease. *Age Ageing*. 2013;**42**:645–649.
6. Eack SM, Greeno CG, Lee BJ. Limitations of the Patient Health Questionnaire in identifying anxiety and depression in community mental health: Many cases are undetected. *Res Soc Work Pract*. 2006;**16**:625–631.
7. Fann JR, Bombardier CH, Dikmen S, et al. Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *J Head Trauma Rehabil*. 2005;**20**:501–511.
8. Fiest KM, Patten SB, Wiebe S, Bulloch AG, Maxwell CJ, Jette N. Validating screening tools for depression in epilepsy. *Epilepsia*. 2014;**55**:1642–1650.

9. Fischer HF, Klug C, Roeper K, et al. Screening for mental disorders in heart failure patients using computer-adaptive tests. *Qual Life Res*. 2014;**23**:1609–1618.
10. Gjerdingen D, Crow S, McGovern P, Miner M, Center B. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann Fam Med*. 2009;**7**:63–70.
11. Gräfe K, Zipfel S, Herzog W, Löwe B. Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. *Diagnostica*. 2004;**50**:171–181.
12. Khamseh ME, Baradaran HR, Javanbakht A, Mirghorbani M, Yadollahi Z, Malek M. Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran. *BMC Psychiatry*. 2011;**11**:61.
13. Kwan Y, Tham WY, Ang A. Validity of the Patient Health Questionnaire-9 (PHQ-9) in the screening of post-stroke depression in a multi-ethnic population. *Biol Psychiatry*. 2012;**71**:141S–141S.
14. 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**:1203–1211.
15. Liu SI, Yeh ZT, Huang HC, et al. Validation of Patient Health Questionnaire for depression screening among primary care patients in Taiwan. *Compr Psychiatry*. 2011;**52**:96–101.
16. McGuire AW, Eastwood JA, Macabasco-O'Connell A, Hays RD, Doering LV. Depression screening: utility of the Patient Health Questionnaire in patients with acute coronary syndrome. *Am J Crit Care*. 2013;**22**:12–19.

17. Osório FL, Vilela Mendes A, Crippa JA, Loureiro SR. Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of Brazilian women in the context of primary health care. *Perspect Psychiatr Care*. 2009;**45**:216–227.
18. Osório FL, Carvalho AC, Fracalossi TA, Crippa JA, Loureiro ES. Are two items sufficient to screen for depression within the hospital context? *Int J Psychiatry Med*. 2012;**44**:141–148.
19. Picardi A, Adler DA, Abeni D, et al. Screening for depressive disorders in patients with skin diseases: a comparison of three screeners. *Acta Derm Venereol*. 2005;**85**:414–419.
20. Richardson TM, He H, Podgorski C, Tu X, Conwell Y. Screening depression aging services clients. *Am J Geriatr Psychiatry*. 2010;**18**:1116–1123.
21. Rooney AG, McNamara S, Mackinnon M, et al. Screening for major depressive disorder in adults with cerebral glioma: an initial validation of 3 self-report instruments. *Neuro-oncology*. 2013;**15**:122–129.
22. Sidebottom AC, Harrison PA, Godecker A, Kim H. Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. *Arch Womens Ment Health*. 2012;**15**:367–374.
23. Simning A, van Wijngaarden E, Fisher SG, Richardson TM, Conwea Y. Mental healthcare need and service utilization in older adults living in public housing. *Am J Geriatr Psychiatry*. 2012;**20**:441–451.
24. Turner A, Hambridge J, White J, et al. Depression screening in stroke: a comparison of alternative measures with the structured diagnostic interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (major depressive episode) as criterion standard. *Stroke*. 2012;**43**:1000–1005.

25. Twist K, Stahl D, Amiel SA, Thomas S, Winkley K, Ismail K. Comparison of depressive symptoms in type 2 diabetes using a two-stage survey design. *Psychosom Medicine*. 2013;**75**:791–797.
26. Vohringer PA, Jimenez MI, Igor MA, et al. Detecting mood disorder in resource-limited primary care settings: comparison of a self-administered screening tool to general practitioner assessment. *J Med Screen*. 2013;**20**:118–124.
27. Williams JR, Hirsch ES, Anderson K, et al. A comparison of nine scales to detect depression in Parkinson disease: which scale to use? *Neurology*. 2012;**78**:998–1006.
28. Wittkampf K, van Ravesteijn H, Baas K, et al. The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *Gen Hosp Psychiatry*. 2009;**31**:451–459.
29. Arroll B, Goodyear-Smith F, Crengle S, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med*. 2010;**8**:348–353.
30. Azah MN, Shah ME, Shaaban J, Bahri IS, Rushidi WM, Jamil YM. Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. *MedPulse*. 2005;**12**:259–263.
31. De Man-van Ginkel JM, Hafsteinsdóttir T, Lindeman E, Burger H, Grobbee D, Schuurmans M. An efficient way to detect poststroke depression by subsequent administration of a 9-item and a 2-item Patient Health Questionnaire. *Stroke*. 2012;**43**:854–856.
32. Delgadillo J, Payne S, Gilbody S, et al. How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. *J Affect Disord*. 2011;**134**:266–271.

33. Gelaye B, Tadesse MG, Williams MA, Fann JR, Vander Stoep A, Zhou XH. Assessing validity of a depression screening instrument in the absence of a gold standard. *Ann Epidemiol*. 2014;**24**:527–531.

34. Hahn D, Reuter K, Harter M. Screening for affective and anxiety disorders in medical patients - comparison of HADS, GHQ-12 and Brief-PHQ. *GMS Psychsoc Med*. 2006;3.

35. Henkel V, Mergl R, Kohnen R, Allgaier AK, Moller HJ, Hegerl U. Use of brief depression screening tools in primary care: consideration of heterogeneity in performance in different patient groups. *Gen Hosp Psychiatr*. 2004;**26**:190–198.

36. Hobfoll SE, Canetti D, Hall BJ, et al. Are community studies of psychological trauma's impact accurate? A study among Jews and Palestinians. *Psychol Assess*. 2011;**23**:599–605.

37. Kiely KM, Butterworth P. Validation of four measures of mental health against depression and generalized anxiety in a community based sample. *Psychiatry Res*. 2014;**225**:291–298.

38. Mohd Sidik S, Arroll B, Goodyear-Smith F. Criterion validity of the PHQ-9 (Malay version) in a primary care clinic in Malaysia. *Med J Malaysia*. 2012;**67**:309–315.

39. Patel V, Araya R, Chowdhary N, et al. Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychol Med*. 2008;**38**:221–228.

40. Pence BW, Gaynes BN, Atashili J, et al. Validity of an interviewer-administered Patient Health Questionnaire-9 to screen for depression in HIV-infected patients in Cameroon. *J Affect Disord*. 2012;**143**:208–213.

41. Razykov I, Hudson M, Baron M, Thombs BD, Canadian Scleroderma Research Group. Utility of the Patient Health Questionnaire-9 to assess suicide risk in patients with systemic sclerosis. *Arth Care Res*. 2013;**65**:753–758.

42. Thombs BD, Ziegelstein RC, Whooley MA. Optimizing detection of major depression among patients with coronary artery disease using the Patient Health Questionnaire: data from the heart and soul study. *J Gen Intern Med*. 2008;**23**:2014–2017.

43. Akena D, Joska J, Obuku EA, Stein DJ. Sensitivity and specificity of clinician administered screening instruments in detecting depression among HIV-positive individuals in Uganda. *AIDS Care*. 2013;**25**:1245–1252.

44. Cholera R, Gaynes BN, Pence BW, et al. Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. *J Affect Disord*. 2014;**167**:160–166.

45. Hides L, Lubman DI, Devlin H, Cotton S, et al. Reliability and validity of the Kessler 10 and Patient Health Questionnaire among injecting drug users. *Aust N Z Psychiatry*. 2007;**41**:166–168.

46. Hyphantis T, Kotsis K, Voulgari PV, Tsifetaki N, Creed F, Drosos AA. Diagnostic accuracy, internal consistency, and convergent validity of the Greek version of the Patient Health Questionnaire 9 in diagnosing depression in rheumatologic disorders. *Arthritis Care Res*. 2011;**63**:1313–1321.

47. Hyphantis T, Kroenke K, Papatheodorou E, et al. Validity of the Greek version of the PHQ 15-item Somatic Symptom Severity Scale in patients with chronic medical conditions and correlations with emergency department use and illness perceptions. *Compr Psychiatry*. 2014;**55**:1950–1959.

48. Inagaki M, Ohtsuki T, Yonemoto N, et al. Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: a cross-sectional study. *Gen Hosp Psychiatry*. 2013;**35**:592–597.

49. Lamers F, Jonkers CC, Bosma H, Penninx BW, Knottnerus JA, van Eijk JT. Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. *J Clin Epidemiol*. 2008;**61**:679–687.

50. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry*. 2008;**8**:46.

51. Muramatsu K, Miyaoka H, Kamijima K, et al. The Patient Health Questionnaire, Japanese version: validity according to the Mini-International Neuropsychiatric Interview-Plus. *Psychol Rep*. 2007;**101**:952–960.

52. Persoons P, Luyckx K, Fischler B. Psychiatric diagnoses in Gastroenterolgy: Validation of a self-report instrument (PRIME-MD Patient Health Questionnaire), epidemiology and recognition. *Gastroenterology*. 2001;**120**:A114–A114.

53. Santos IS, Tavares BF, Munhoz TN, et al. [Sensitivity and specificity of the Patient Health Questionnaire-9 (PHQ-9) among adults from the general population]. *Cad Saude Publica*. 2013;**29**:1533–1543.

54. Stafford L, Berk M, Jackson HJ. Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. *Gen Hosp Psychiatry*. 2007;**29**:417–424.

55. Sung SC, Low CC, Fung DS, Chan YH. Screening for major and minor depression in a multiethnic sample of Asian primary care patients: a comparison of the nine-item Patient Health Questionnaire (PHQ-9) and the 16-item Quick Inventory of Depressive Symptomatology - Self-Report (QIDS-SR16). *Asia Pac Psychiatry*. 2013;**5**:249–258.

56. van Steenbergen-Weijenburg KM, de Vroege L, Ploeger RR, et al. Validation of the PHQ-9 as a screening instrument for depression in diabetes patients in specialized outpatient clinics. *BMC Health Serv Res*. 2010;**10**:235.

57. Zhang Y, Ting R, Lam M, et al. Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong Chinese subjects with type 2 diabetes. *J Affect Disord*. 2013;**151**:660–666.

58. Becker S, Al Zaid K, Al Faris E. Screening for somatization and depression in Saudi Arabia: a validation study of the PHQ in primary care. *Int J Psychiatry Med*. 2002;**32**:271–283.

59. Chen S, Fang Y, Chiu H, Fan H, Jin T, Conwell Y. Validation of the nine-item Patient Health Questionnaire to screen for major depression in a Chinese primary care population. *Asia Pac Psychiatry*. 2013;**5**:61–68.

60. Chen S, Conwell Y, Vanorden K, et al. Prevalence and natural course of late-life depression in China primary care: a population based study from an urban community. *J Affect Disord*. 2012;**141**:86–93.

61. Lai BP, Tang AK, Lee DT, Yip AS, Chung TK. Detecting postnatal depression in Chinese men: a comparison of three instruments. *Psychiatry Res*. 2010;**180**:80–85.

62. Navines R, Castellvi P, Moreno-Espana J, et al. Depressive and anxiety disorders in chronic hepatitis C patients: reliability and validity of the Patient Health Questionnaire. *J Affect Disord*. 2012;**138**:343–351.

63. Phelan E, Williams B, Meeker K, et al. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract*. 2010;**11**:63.

64. Thompson AW, Liu H, Hays RD, et al. Diagnostic accuracy and agreement across three depression assessment measures for Parkinson's disease. *Parkinsonism Relat Disord*. 2011;**17**:40–45.

65. Watnick S, Wang PL, Demadura T, Ganzini L. Validation of 2 depression screening tools in dialysis patients. *Am J Kid Dis*. 2005;**46**:919–924.

66. Al-Ghafri G, Al-Sinawi H, Al-Muniri A, et al. Prevalence of depressive symptoms as elicited by Patient Health Questionnaire (PHQ-9) among medical trainees in Oman. *Asian J Psychiatr*. 2014;**8**:59–62.

67. Haddad M, Walters P, Phillips R, et al. Detecting depression in patients with coronary heart disease: a diagnostic evaluation of the PHQ-9 and HADS-D in primary care, findings from the UPBEAT-UK study. *PLoS ONE*. 2013;**8**:e78493.

68. Persoons P, Luyckx K, Desloovere C, Vandenberghe J, Fischler B. Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology. *Gen Hosp Psychiatry*. 2003;**25**:316–323.

69. Rathore JS, Jehi LE, Fan Y, et al. Validation of the Patient Health Questionnaire-9 (PHQ-9) for depression screening in adults with epilepsy. *Epilepsy Behav*. 2014;**37**:215–220.

70. Scott JD, Wang CC, Coppel E, Lau A, Veitengruber J, Roy-Byrne P. Diagnosis of depression in former injection drug users with chronic hepatitis C. *J Clin Gastroenterol*. 2011;**45**:462–467.

71. Wang W, Bian Q, Zhao Y, et al. Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population. *Gen Hosp Psychiatry*. 2014;**36**:539–544.