External Validation of a Shortened Screening Tool Using Individual Participant Data Meta-Analysis: a Case Study of the Patient Health Questionnaire-Dep-4

Daphna Harel[1,2]; Brooke Levis[3]; Ying Sun[4], Felix Fischer[5], John P.A. Ioannidis[6-9], Pim Cuijpers[10], Scott B. Patten[11], Roy C. Ziegelstein[12], Sarah Markham[13], Andrea Benedetti[14-16], Brett D. Thombs[4,14,15,17-20], and the DEPRESsion Screening Data (DEPRESSD) PHQ Collaboration[21]

[1]Department of Applied Statistics, Social Science, and Humanities, New York University, United States

[2]Center for the Promotion of Research at the Intersection of Information, Society, and Methodology, New York University, United States

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

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

[5]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

[6]Department of Medicine, Stanford University, Stanford, California, USA

[7]Department of Epidemiology and Population Health, Stanford University, Stanford, California, USA

[8]Department of Biomedical Data Science, Stanford University, Stanford, California, USA

[9]Department of Statistics, Stanford University, Stanford, California, USA

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

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

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

[13]Department of Biostatistics and Health Informatics, King's College London, London, UK

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

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

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

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

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

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

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

[21]The DEPRESSD PHQ Collaboration: Chen He, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Yin Wu, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Ankur Krishnan, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Parash Mani Bhandari, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Dipika Neupane, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Zelalem Negeri, Lady Davis Institute for Medical

Research, Jewish General Hospital, Montréal, Québec, Canada; Mahrukh Imran, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Danielle B. Rice, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Kira E. Riehm, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Marleine Azar, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Alexander W. Levis, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Jill Boruff, Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montréal, Québec, Canada; Simon Gilbody, Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK; Lorie A. Kloda, Library, Concordia University, Montréal, Québec, Canada; Dagmar Amtmann, Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA; Liat Ayalon, Louis and Gabi Weisfeld School of Social Work, Bar Ilan University, Ramat Gan, Israel; Hamid R. Baradaran, Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; Anna Beraldi, Kbo-Lech-Mangfall-Klinik Garmisch-Partenkirchen, Klinik für Psychiatrie, Psychotherapie & Psychosomatik, Lehrkrankenhaus der Technischen Universität München, Munich, Germany; Charles N. Bernstein, University of Manitoba IBD Clinical and Research Centre, Winnipeg, Manitoba, Canada; Arvin Bhana, Centre for Rural Health, School of Nursing and Public Health, College of Health Sciences, University of KwaZulu-Natal; Ryna Imma Buji, Department of Psychiatry, Hospital Mesra Bukit Padang, Sabah, Malaysia; Marcos H. Chagas, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Juliana C. N. Chan, Department of Medicine and Therapeutics, Hong Kong Institute of Diabetes and Obesity and Li Ka Shing

Institute of Health Science, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong SAR, China; Lai Fong Chan, Department of Psychiatry, National University of Malaysia, Kuala Lumpur, Malaysia; Dixon Chibanda, Department of Community Medicine, University of Zimbabwe, Harare, Zimbabwe; Aaron Conway, Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, Toronto, Canada; Federico M. Daray, Institute of Pharmacology, School of Medicine, University of Buenos Aires, Argentina; Janneke M. de Man-van Ginkel, Julius Center for Health Sciences and Primary Care, Department of Nursing Science, University Medical Center Utrecht – University Utrecht, Utrecht, the Netherlands; Crisanto Diez-Quevedo, Servei de Psiquiatria, Hospital Germans Trias i Pujol, Badalona, Spain; Sally Field, Perinatal Mental Health Project, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Jane R. W. Fisher, Global and Women's Health, Public Health and Preventive Medicine, Monash University; Daniel Fung, Department of Developmental Psychiatry, Institute of Mental Health, Singapore; Emily C. Garman, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Bizu Gelaye, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; Leila Gholizadeh, Faculty of Health, University of Technology Sydney, Sydney, Australia; Lorna J. Gibson, International Statistics and Epidemiology Group, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK; Eric P. Green, Duke Global Health Institute, Duke University, Durham, North Carolina, USA; Brian J. Hall, New York University Shanghai, Shanghai, People's Republic of China; Liisa Hantsoo, Department of Psychiatry & Behavioral Sciences , The Johns Hopkins University School of Medicine, Baltimore, Maryland; Emily E. Haroz, Center For American Indian Health, Department of International Health, Johns

Hopkins Bloomberg School of Public Health; Martin Härter, Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Ulrich Hegerl, Department of Psychiatry, Psychosomatics and Psychotherapy, Goethe-Universität Frankfurt, Germany; Leanne Hides, School of Psychology, University of Queensland, Brisbane, Queensland, Australia; Stevan E. Hobfoll, STAR-Stress, Anxiety and Resilience Consultants, Chicago, Illinois, USA; Simone Honikman, Perinatal Mental Health Project, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Marie Hudson, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Thomas Hyphantis, Department of Psychiatry, Faculty of Medicine, School of Health Sciences, University of Ioannina, Greece; Masatoshi Inagaki, Department of Psychiatry, Faculty of Medicine, Shimane University, Shimane, Japan; Hong Jin Jeon, Department of Psychiatry, Depression Center, Samsung Medical Center, Sungkyunkwan University School of Medicine; Nathalie Jetté, Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, New York, USA; Mohammad E. Khamseh, Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran; Sebastian Köhler, Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, Netherlands; Brandon A. Kohrt, Department of Psychiatry and Behavioral Sciences, The George Washington University, Washington, DC, USA; Yunxin Kwan, Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore; Femke Lamers, Department of Psychiatry, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit, Amsterdam, the Netherlands; Maria Asunción Lara, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz. San Lorenzo Huipulco, Tlalpan, México D. F. Mexico; Holly F. Levin-Aspenson, Department of Psychology, University of Notre Dame, Notre

Dame, Indiana, USA; Shen-Ing Liu, Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore; Manote Lotrakul, Department of Psychiatry, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; Sonia R. Loureiro, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Bernd Löwe, Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Nagendra P. Luitel, Research Department, TPO Nepal, Kathmandu, Nepal; Crick Lund, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town; Ruth Ann Marrie, Departments of Medicine and Community Health Sciences, Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada; Brian P. Marx, National Center for PTSD at VA Boston Healthcare System, Boston, MA, USA; Sherina Mohd Sidik, Department of Psychiatry, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; Tiago N. Munhoz, Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil; Kumiko Muramatsu, Department of Clinical Psychology, Graduate School of Niigata Seiryo University, Niigata, Japan; Juliet E. M. Nakku, Butabika National Referral Teaching Hospital, Kampala, Uganda; Laura Navarrete, Department of Epidemiology and Psychosocial Research, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Ciudad de México, México; Flávia L. Osório, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil; Philippe Persoons, Department of Psycho-Pedagogic Psychiatry, Healthcare Group Sint-Kamillus, Broeders van Liefde, Bierbeek, Belgium; Angelo Picardi, Centre for Behavioural Sciences and Mental Health, Italian National Institute of Health, Rome, Italy; Stephanie L. Pugh, NRG Oncology Statistics and Data

Management Center, Philadelphia, PA, USA; Terence J. Quinn, Institute of Cardiovascular & Medical Sciences, University of Glasgow, Glasgow, Scotland; Elmars Rancans, Department of Psychiatry and Narcology, Riga Stradins University, Latvia; Sujit D. Rathod, Department of Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom; Katrin Reuter, Group Practice for Psychotherapy and Psycho-oncology, Freiburg, Germany; Heather J. Rowe, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; Iná S. Santos, Post-graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, RS, Brazil; Miranda T. Schram, Department of Internal Medicine, Maastricht University Medical Center, Maastricht, The Netherlands; Juwita Shaaban, Department of Family Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia; Eileen H. Shinn, Department of Behavioral Science, University of Texas M. D. Anderson Cancer Center, Houston, Texas, USA; Lena Spangenberg, Department of Medical Psychology and Medical Sociology, University of Leipzig, Germany; Lesley Stafford, Centre for Women's Mental Health, Royal Women's Hospital, Parkville, Australia; Sharon C. Sung, Programme in Health Services & Systems Research, Duke-NUS Medical School, Singapore; Keiko Suzuki, Department of General Medicine, Asahikawa University Hospital, Asahikawa, Hokkaido, Japan; Pei Lin Lynnette Tan, Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore; Martin Taylor-Rowan, Institute of Cardiovascular and Medical Science, University of Glasgow, Glasgow, Scotland; Thach D. Tran, Global and Women's Health, Public Health and Preventive Medicine, Monash University; Christina M. van der Feltz-Cornelis, Department of Health Sciences, HYMS, University of York, York, UK; Thandi van Heyningen, Division of Epidemiology & Biostatistics, School of Public Health & Family Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa; Henk C. van Weert,

Department of General Practice, Institute Public Health, Amsterdam Universities Medical Centers, Amsterdam, the Netherlands; Lynne I. Wagner, Department of Social Sciences and Health Policy, Wake Forest School of Medicine, Wake Forest University, Winston-Salem, North Carolina, USA; Jian Li Wang, University of Ottawa Institute of Mental Health Research; David Watson, Dept. of Psychology, University of Notre Dame; Karen Wynter, School of Nursing and Midwifery, Deakin University, Melbourne, Australia; Mitsuhiko Yamada, Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Ogawa-Higashi, Kodaira, Tokyo, Japan; Qing Zhi Zeng, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China; Yuying Zhang, Department of Medicine and Therapeutics, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China.

Corresponding author

Daphna Harel, PhD; 246 Greene Street, 3rd floor, New York, NY, 10003.

daphna.harel@nyu.edu, 212-992-6701.

Abstract

Shortened versions of self-reported questionnaires may be used to reduce respondent burden. When shortened screening tools are used, it is desirable to maintain equivalent diagnostic accuracy to full-length forms. This manuscript presents a case study that illustrates how external data and individual participant data meta-analysis can be used to assess the equivalence in diagnostic accuracy between a shortened and full-length form. This case study compares the Patient Health Questionnaire-9 (PHQ-9) and a 4-item shortened version (PHQ-Dep-4) that was previously developed using optimal test assembly methods. Using a large database of 75 primary studies (34,698 participants, 3,392 major depression cases), we evaluated whether the PHQ-Dep-4 cutoff of ≥ 4 maintained equivalent diagnostic accuracy to a PHQ-9 cutoff of ≥ 10. Using this external validation dataset, a PHQ-Dep-4 cutoff of ≥ 4 maximized the sum of sensitivity and specificity, with a sensitivity of 0.88 (95% CI 0.81, 0.93), 0.68 (95% CI 0.56, 0.78), and 0.80 (95% CI 0.73, 0.85) for the semi-structured, fully structured, and MINI reference standard categories, respectively, and a specificity of 0.79 (95% CI 0.74, 0.83), 0.85 (95% CI 0.78, 0.90), and 0.83 (95% CI 0.80, 0.86) for the semi-structured, fully structured, and MINI reference standard categories, respectively. While equivalence with a PHQ-9 cutoff of ≥ 10 was not established, we found the sensitivity of the PHQ-Dep-4 to be non-inferior to that of the PHQ-9, and the specificity of the PHQ-Dep-4 to be marginally smaller than the PHQ-9.

Keywords: Optimal Test Assembly; Sensitivity; Specificity; Equivalence Testing; Self-report questionnaire

Highlights

1) Optimal Test Assembly is a reproducible and replicable method to create shorter forms and reduce burden on respondents
2) This manuscript is the first paper to externally validate a measure developed through optimal test assembly methods
3) In our validation of the Patient Health Questionnaire 4-item shortened form, we found that the same cutoff maximized diagnostic accuracy
4) We found that sensitivity was non-inferior to that of the full-length form, but the specificity was slightly reduced.

INTRODUCTION

Self-reported symptom measures are used to assess mental health symptoms and may also be used to screen for mental disorders. However, in clinical practice and research, individuals may be asked to complete several measures, each with multiple items or domains, which can be demanding on their time, and sensitive items, such as asking about suicidal ideation, may be emotionally burdensome [1]–[4]. Long measures can result in poor data quality and high amounts of missing data. Thus, shortened forms that do not significantly reduce diagnostic accuracy can provide meaningful data while reducing respondent burden and potentially increasing data quality.

The Patient Health Questionnaire-9 (PHQ-9) is a 9-item, self-report questionnaire that measures depressive symptoms [5]–[7]. Scores on each item on the PHQ-9 range reflect symptoms in the last 2 weeks and range from 0 (“not at all”) to 3 (“every day”). Scores range from 0 to 27 with higher scores indicating higher levels of depressive symptomatology.

An individual participant data meta-analysis (IPDMA) on the accuracy of the PHQ-9 to screen for major depression was conducted on 29 studies with a semi-structured diagnostic interview as the reference standard (6,725 participants, 924 major depression cases). This study found that the standard and most commonly used for the PHQ-9, cutoff threshold of ≥ 10, maximized the combination of sensitivity (0.88, 95% CI 0.83, 0.92) and specificity (0.85, 95% CI 0.82, 0.88) [8].

Using a subset of data from the IPDMA, a previous study developed a 4-item shortened form of the PHQ-9, known as the PHQ-Dep-4, through optimal test assembly (OTA) methods. As with the PHQ-9, scores on each item of the PHQ-Dep-4 reflect symptoms in the last 2 weeks

and range from 0 (“not at all”) to 3 (“every day”). PHQ-Dep-4 scores range from 0 to 12 with higher scores indicating higher levels of depressive symptomatology.

The initial development study used 20 primary studies (7,850 participants, 863 major depression cases), which we refer to as the development sample, that administered the English version of the PHQ-9 and used a validated semi-structured or fully structured diagnostic interview (Mini International Neuropsychiatric Interview [MINI] excluded) to classify major depression. The PHQ-Dep-4 includes items 1, 2, 6, and 8 from the PHQ-9, representing depressed mood, loss of interest/pleasure, low self-esteem/guilt and psychomotor agitation [9]. OTA is a mixed-integer programming procedure that uses an estimated item response theory model to select the subset of items that best satisfies pre-specified constraints. In the case of the PHQ-Dep-4 development study, there were pre-specified constraints on the concurrent validity, reliability, and equivalency of diagnostic accuracy of the shortened form with the full-length form [10]. Although more commonly used in the development of high-stakes educational tests [11], recent studies have demonstrated that OTA can be used to develop shortened versions of patient-reported outcome measures [9], [12]–[17]. This procedure was shown in a simulation study to be replicable and reproducible, and produce shortened forms of minimal length with limited loss of information [14].

A cutoff of ≥ 4 on the PHQ-Dep-4 was found to perform equivalently to the PHQ-9 cutoff ≥ 10 in the development sample. However, accuracy of the PHQ-Dep-4 has not been externally validated outside of the development sample. It is therefore necessary to investigate whether a cutoff of ≥ 4 on the PHQ-Dep-4 continues to maintain equivalent diagnostic accuracy to the PHQ-9 cutoff ≥ 10. Conducting an external validation of this cutoff allows for the assessment of whether this cutoff was specific to the development dataset or generalizable to

other studies or applications in the future. In particular, the development of the PHQ-Dep-4 was based on comparing properties of the full-length form to a set of candidate shortened forms in the development sample, and thus is susceptible to issues of overfitting or a lack of generalizability. By conducting an external validation, it is possible to see whether the equivalence in accuracy of the PHQ-Dep-4 to the PHQ-9 can be confirmed in an independent dataset.

The objective of the present study was to use data from a unique set of studies that administered the PHQ-9 as well as a validated semi-structured or fully structured diagnostic interview for major depression to validate the diagnostic accuracy of the previously developed PHQ-Dep-4. Specifically, we (1) estimated accuracy for all possible PHQ-Dep-4 cutoffs (i.e., ≥ 1 to ≥ 12), and (2) tested equivalency in accuracy for each PHQ-Dep-4 cutoff to that of a PHQ-9 cutoff of ≥ 10, with the comparison of the PHQ-Dep-4 cutoff of ≥ 4 considered the primary comparison.

METHODS

The present validation study used data synthesized from an updated IPDMA of the screening accuracy of the PHQ-9 for major depression [8], [18], excluding datasets that were included in the original PHQ-Dep-4 development project [9]. The present validation study included studies conducted in any language and using any validated semi-structured or fully structured diagnostic interview (MINI included). The main IPDMA was registered in PROSPERO (CRD42014010673) and a protocol was published [19]. The present analysis was not part of the protocol for the main IPDMA, but a separate protocol was developed and posted prior to initiation at https://osf.io/xy2b8/.

The Main IPDMA Database

Study selection

In the main IPDMA, datasets from articles in any language were eligible for inclusion if (1) they included PHQ-9 scores; (2) they included diagnostic classifications for current Major Depressive Episode (MDE) or Major Depressive Disorder (MDD) based on Diagnostic and Statistical Manual of Mental Disorders (DSM) [20]–[23], or International Classification of Diseases (ICD) [24] criteria, using a validated semi-structured or fully structured interview; (3) the PHQ-9 and diagnostic interview were administered within two weeks of each other, since diagnostic criteria for major depression are for symptoms in the last two weeks; (4) participants were ≥ 18 years and not recruited from youth or school-based settings; and (5) participants were not recruited from psychiatric settings or because they were identified as having symptoms of depression, since screening is done to identify unrecognized cases. Datasets where not all participants were eligible were included if primary data allowed selection of eligible participants.

Database sources and search strategy

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations via Ovid; PsycINFO; and Web of Science from January 1, 2000 to May 9, 2018 using a peer-reviewed search strategy (eMethods1) [25]. The search was limited to the year 2000 onwards because the PHQ-9 was first published in 2001 [7]. We also reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After deduplication, remaining citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada) for processing review results.

Two investigators independently reviewed titles and abstracts for eligibility. If either investigator deemed a study potentially eligible, full-text review was done by two investigators, independently, with disagreements resolved by consensus, consulting a third investigator when

necessary. Translators were consulted for languages other than those for which team members were fluent.

Data contribution and synthesis

Authors of eligible datasets were invited to contribute de-identified primary data, including PHQ-9 scores and major depression status. We emailed corresponding authors of eligible primary studies at least three times, as necessary, with at least two weeks between each email. If we did not receive a response, we emailed co-authors and attempted to contact corresponding authors by phone.

Individual participant data were converted to a standard format and synthesized into a single dataset with study-level data. We compared published participant characteristics and diagnostic accuracy results with results from raw datasets and resolved any discrepancies in consultation with the original investigators.

To define major depression, we considered MDD or MDE based on the DSM or ICD. If more than one was reported, we prioritized MDE over MDD, since screening would attempt to detect depressive episodes and further interview would determine if the episode were related to MDD, bipolar disorder, or persistent depressive disorder. When both were present, we prioritized DSM over ICD, because DSM is more commonly used in existing studies.

Data Used in the Present Analyses

To consider an independent data source for this validation, we excluded the 20 studies that were included in the original PHQ-Dep-4 development project. We note that these 20 studies were originally used in the development paper because of their availability at the time that study was conducted, rather than a deliberate splitting of the sample. In addition, to be able to calculate PHQ-Dep-4 scores, we excluded studies and participants without item-level PHQ-9 data.

Statistical Analyses

Using the item-level PHQ-9 data, we calculated PHQ-Dep-4 scores by summing the item scores from PHQ-9 items 1 (loss of interest), 2 (depressed mood), 6 (feeling like a failure), and 8 (physical movement). We then conducted two sets of analyses.

To assess diagnostic accuracy, we estimated sensitivity and specificity. Sensitivity, the true positive rate, refers to the probability of scoring above the cutoff in question given that the participant was classified with MDE or MDD based on DSM or ICD criteria using a validated semi-structured or fully structured interview. Specificity, the true negative rate, refers to the probability of scoring below the cutoff in question given that the participant was classified with MDE or MDD based on DSM or ICD criteria using a validated semi-structured or fully structured interview.

First, we estimated sensitivity and specificity for all possible PHQ-Dep-4 cutoffs (i.e., ≥ 1 to ≥ 12), as well as the standard PHQ-9 cutoff score of ≥ 10, which maximizes sensitivity + specificity [8], [18]. For each PHQ-Dep-4 cutoff, separately, and for a PHQ-9 cutoff of ≥ 10, we fit bivariate random-effects models using adaptive Gauss-Hermite quadrature with one quadrature point [26]. This is a 2-stage meta-analytic approach that synthesizes sensitivity and specificity simultaneously and accounts for the correlation between them, as well as for precision of estimates within studies. For each analysis, this model provided estimates of pooled sensitivity and specificity.

The formulation of the model can be expressed as the following. Let $y_{s,i}^{(0)}$ be the dichotomous outcome of the screening test (PHQ-9 or PHQ-Dep-4) for the *i*-th participant in the *s*-th primary study who does not have a true depression diagnosis. Therefore, $y_{s,i}^{(0)}$ is equal to one when the participant has a high score on the screening test and zero when the participant has a low score on the screening test. Similarly, let $y_{s,i}^{(1)}$ be the dichotomous outcome of the screening

test for the *i*-th participant of the *s*-th primary study who does have a true depression diagnosis. The model is formulated as:

$$y_{s,i}^{(0)} \sim Bernoulli(p_{s,i}^{(0)})$$

$$logit\left(p_{s,i}^{(0)}\right) = \mu_s^{(0)} = \mu^{(0)} + u_s^{(0)}$$

$$y_{s,i}^{(1)} \sim Bernoulli(p_{s,i}^{(1)})$$

$$logit\left(p_{s,i}^{(1)}\right) = \mu_s^{(1)} = \mu^{(1)} + u_s^{(1)}$$

$$\boldsymbol{u}_s = \begin{pmatrix} u_s^{(0)} \\ u_s^{(1)} \end{pmatrix} \sim N(0, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \tau_0^2 & \tau_0\tau_1\rho_\tau \\ \tau_0\tau_1\rho_\tau & \tau_1^2 \end{pmatrix}$$

In this case, the false positive rate (FPR), which is equal to 1 – specificity, and the true positive rate (TPR), which is the sensitivity, can be estimated for the pooled logit(FPR) and logit(TPR) through $\hat{\mu}^{(0)}$and $\hat{\mu}^{(1)}$, respectively. $\hat{\tau}^{(0)}$and $\hat{\tau}^{(1)}$estimates the between-study variance of the logit-transformed parameters, and $\widehat{\rho_\tau}$ is the estimated correlation.

For these analyses, we modeled sensitivity and specificity separately among studies that used each reference standard category (semi-structured, fully structured, or MINI) as well as pooled together. We present accuracy results for the PHQ-Dep-4 separately by reference standard type because previous studies have found that there are important differences in the design and performance of different types of diagnostic interviews used as reference standards [27]–[30], and that PHQ-9 sensitivity and specificity vary across different reference standards [8], [18]. For each reference standard category, we constructed an empirical receiver operating characteristic (ROC) plot for the PHQ-Dep-4 based on pooled sensitivity and specificity

estimates from each cutoff. Separately, we marked the point in ROC-space for a PHQ-9 cutoff of ≥ 10.

Second, we tested the equivalence of the PHQ-Dep-4 and PHQ-9. The comparison of the PHQ-Dep-4 cutoff of ≥ 4 to the PHQ-9 cutoff of ≥ 10 was considered as our primary analysis. For these analyses, we pooled reference standard categories together, because although PHQ-9 and PHQ-Dep-4 sensitivity and specificity may differ by reference standard category, we did not believe that *differences* in sensitivity and specificity between PHQ-Dep-4 cutoffs and a PHQ-9 cutoff of ≥ 10 would vary by reference standard category, since each primary study compared the PHQ-Dep-4 and PHQ-9 to the same reference standard. By pooling, we increase power and therefore reduce the risk of an ambiguous outcome in the analysis. In line with this, a previous comparison of the PHQ-8 and PHQ-9 found that although accuracy differed across reference standard categories, differences in accuracy across the forms were similar across reference standard categories [31]. We estimated the crude differences in sensitivity and specificity between each PHQ-Dep-4 cutoff and a PHQ-9 cutoff of ≥ 10 and constructed confidence intervals (CI) for differences via the cluster bootstrap approach [32], [33], resampling at study and subject levels with replacement. For each comparison, we ran 1000 iterations of the bootstrap. These CIs allowed us to test whether the sensitivity and specificity of each PHQ-Dep-4 cutoff are equivalent to that of the PHQ-9 based on a pre-specified minimally important difference of $\delta = 0.05$ [34], as has been done in previous studies [9], [13], [31]. That is, for each cutoff, for differences in sensitivity and specificity separately, we would consider the null hypothesis that there are differences large enough to be important and test that against the alternative hypothesis that there are no meaningful differences. If the entire CI is included within the interval of − 0.05 to + 0.05, we would reject the null hypothesis and conclude that

equivalence is present. If the entire CI is outside of the interval, we would conclude that the accuracies are not equivalent. If the CIs cross the interval of − 0.05 to + 0.05, findings would be deemed ambiguous, and the equivalence would be found to be indeterminate. Lastly, we determined which PHQ-Dep-4 cutoff showed the smallest overall sum of absolute differences in accuracy (i.e. in sensitivity and in specificity) compared to PHQ-9 ≥ 10.

All analyses were conducted in R (R version R 3.4.1 [35], RStudio version 1.0.143) using the *glmer* function within the *lme4* package [36]. All R code used to run the analysis is included in the supplementary materials, however due to data sharing agreements, the raw data is not available.

Ethics

As this study involves secondary analysis of de-identified previously collected data, the Research Ethics Committee of the Jewish General Hospital determined that it did not require research ethics approval. However, for each included dataset, we confirmed that the original study received ethics approval and that all participants provided informed consent.

RESULTS

Search Results and Dataset Inclusion

Figure 1 illustrates the study flow diagram. Of 9,670 unique titles and abstracts identified from database searches, 9,199 were excluded at the title and abstract review stage and 297 after full-text review. After removing duplicate samples, adding unpublished studies contributed by authors, excluding studies that did not have item level data or were included in the PHQ-Dep-4 development paper, there were 75 eligible datasets (N participants = 34,698; N major depression = 3,392 [prevalence 10%]) that contributed data for our analysis.

Of the 75 included studies, 29 (7,719 participants; 923 major depression cases) used a semi-structured interview as the reference standard, 15 (12,109 participants; 873 cases) used a fully structured interview (other than the MINI), and 31 (14,870 participants; 1,596 cases) used the MINI. The Structured Clinical Interview for the DSM (SCID) was the most commonly used semi-structured interview (28 of 29 studies) and the Composite International Diagnostic Interview (CIDI) the most commonly used fully structured interview (14 of 15 studies). See Supplementary Table 1a-c for characteristics of included primary studies, eligible excluded primary studies, and the 20 studies included in the PHQ-Dep-4 development paper only. Table 1 presents participant-level descriptive statistics for the sample used in the present study.

Validation Results

Figure 2 shows receiver-operating curves for each reference standard category as well as the PHQ-9 cutoff score of ≥ 10. Table 2 shows estimated sensitivity and specificity for PHQ-Dep-4 cutoffs (≥ 1 to ≥ 12), as well as the standard and optimal PHQ-9 cutoff score of ≥ 10. For a PHQ-Dep-4 cutoff of ≥ 4, sensitivity was 0.88 (95% CI 0.81, 0.93), 0.68 (95% CI 0.56, 0.78), and 0.80 (95% CI 0.73, 0.85) for the semi-structured, fully structured, and MINI reference standard categories, respectively, as compared to 0.88 (0.81, 0.93), 0.64 (0.50, 0.76), and 0.73 (0.66, 0.79) for the PHQ-9 cutoff of ≥ 10, respectively. Similarly, for a PHQ-Dep-4 cutoff of ≥ 4, specificity was 0.79 (95% CI 0.74, 0.83), 0.85 (95% CI 0.78, 0.90), and 0.83 (95% CI 0.80, 0.86) for the semi-structured, fully structured, and MINI reference standard categories, respectively, as compared to 0.85 (0.80, 0.88), 0.89 (0.83, 0.93), and 0.89 (0.86, 0.91) for the PHQ-9 cutoff of ≥ 10, respectively. Figure 2 shows the ROC plots for each reference standard category.

Table 3 shows the results of the tests of equivalence of the PHQ-Dep-4 and PHQ-9 pooled across all reference standard categories. A PHQ-Dep-4 cutoff of ≥ 4 showed the smallest

overall sum of absolute differences in accuracy with PHQ-9 ≥ 10, with a difference in sensitivity of 0.03 (95% CI 0.00, 0.06) and a difference in specificity of -0.05 (95% CI -0.07, -0.04). These findings were ambiguous, as the CIs for both sensitivity and specificity crossed the interval of -0.05 to +0.05. No other PHQ-Dep-4 cutoff indicated equivalency for both sensitivity and specificity. The next closest PHQ-Dep-4 cutoff to PHQ-9 ≥ 10 was a PHQ-Dep-4 cutoff of ≥ 5, with a difference in sensitivity of -0.07 (95% CI -0.11, -0.05) and a difference in specificity of 0.02 (95% CI 0.01, 0.03).

DISCUSSION

This study used data from 75 primary studies to assess whether a previously determined PHQ-Dep-4 cutoff of ≥ 4, which was equivalent to a PHQ-9 cutoff of ≥ 10 in a development sample, would also be equivalent in a validation sample. While a PHQ-Dep-4 cutoff of ≥ 4 showed the best performance among all possible PHQ-Dep-4 cutoffs compared to the PHQ-9 cutoff of ≥ 10, the equivalence results were ambiguous, and we were unable to conclude that its specificity was equivalent to that of the PHQ-9 cutoff of ≥ 10.

We found that compared to the standard and optimal PHQ-9 cutoff of ≥ 10, a PHQ-Dep- cutoff of ≥ 4 had slightly greater sensitivity and slightly reduced specificity. The next best PHQ-Dep-cutoff of ≥ 5 had slightly greater specificity and slightly reduced sensitivity. In clinical settings, use of shortened forms such as the PHQ-Dep-4 offers the advantage of reducing respondent burden. While our study assessed the sum of sensitivity and specificity, this does not necessarily reflect local concerns such as the capacity for conducting further assessments, nor does it necessarily maximize the likelihood of patient benefits or minimize costs and harms. We note that clinicians and researchers can choose different cut-offs based on local priorities and resources using the information provided in Tables 2 and 3.

While a strength of this analysis is the large number of primary studies included in the dataset, these primary studies spanned a large number of languages. This can cause concern for differential item functioning (DIF). The items for the PHQ-Dep-4 were not selected with regards to considerations of DIF. However, studies of DIF with the PHQ-9 have shown that it performs equivalently or with minimal impact of DIF across multiple languages [37]–[39]. We note that future research may wish to specifically investigate the impact of DIF for the PHQ-Dep-4 in comparison to the PHQ-9.

The development study tested non-inferiority rather than equivalency. The development study found a difference in sensitivity of +0.03, and a difference in specificity of -0.03 between the two forms [9]. The present study found differences of +0.03 and -0.05, respectively. While equivalency is therefore not established, the findings in the present study were not substantively different from the development study.

While it is not clear that the PHQ-Dep-4 performs equivalently to the PHQ-9 for specificity, clinicians screening for depression may opt to use the PHQ-Dep-4 with the understanding that depending on the cutoff used, specificity might be slightly reduced compared to the full PHQ-9 at cutoff of ≥ 10. Furthermore, clinicians should be aware that while the full PHQ-9 aligns with the nine DSM symptoms for major depression, not all PHQ-9 items may be relevant to individual presentations of a given mental disorder, and the PHQ-Dep-4 includes only a pre-specified subset of four items (1, 2, 6, and 8), thus not necessarily capturing the specific symptoms of a given patient.

There are several reasons that may explain why equivalence could not be concluded. First, although the overall sample size and number of studies used in this analysis was large, it could be that the study was underpowered, due to the design effect associated with the clustering

within studies. As we do not know of methods for calculating power to establish equivalency in accuracy based on sensitivity and specificity difference for a subset of items compared to the total set, it was not possible to determine the necessary sample size needed *a priori*. Furthermore, we also did not split the data by reference standard category and conduct separate analyses. Second, we found that sensitivity in the shortened form was improved as compared to the full-length form. However, the specificity of the shortened form was lower than that of the full-length form, resulting in the inability to conclude equivalence between the two forms.

There are several other possible limitations of this study. First, for the collection of data for the full IPDMA, we were unable to obtain data from 27 eligible studies. Of the studies that provided data, five were excluded because they did not include item-level scores necessary to calculate PHQ-Dep-4, and we excluded another 20 studies from the development dataset to provide us with a set of external validation data. With the final available dataset, we were unable to investigate equivalence in specific patient populations as that would have required splitting the data even further. Second, for our first set of analyses (estimating PHQ-Dep-4 accuracy at all cutoffs), primary studies were categorized based on the diagnostic interview used, but interviewers may not have always administered the interviews as intended, which could have influenced results. This study only compared the PHQ-Dep-4 to a PHQ-9 cutoff of ≥ 10 because, although some primary studies have found other preferred cutoffs, large IPDMAs have concluded that cutoff ≥ 10 maximizes the sum of sensitivity and specificity [8], [18]. Lastly, this study evaluated the items included in the PHQ-Dep-4 as previously developed and did not re-develop the shortened form. It could be that a different set of items, creating either a different form of length 4 or a potentially shorter or longer form, would result in equivalent sensitivity and specificity to the full PHQ-9.

CONCLUSION

In conclusion, this was the first study to our knowledge to externally validate the results of shortening a self-report questionnaire through the OTA method using individual participant level data. We found that the previously suggested cutoff of ≥ 4 for the PHQ-Dep-4 remained the preferred cutoff, but the specificity of the shortened form did not meet equivalency to the full PHQ-9 cutoff of ≥ 10. Clinicians may consider screening with the PHQ-Dep-4 to reduce respondent burden, but should be aware that in doing so, specificity may be slightly compromised compared to the full PHQ-9.

Contributions:

DH, BLevis, JPAI, PC, SBP, RCZ, ABenedetti, and BDT were responsible for the study conception and design. SM contributed as a patient partner knowledge user. FF contributed an included dataset. BLevis, YS, and BDT contributed to data extraction, coding, evaluation of included studies, and data synthesis. DH, BLevis, FF, ABenedetti, and BDT contributed to data analysis and interpretation. DH, BLevis, YS, ABenedetti, and BDT drafted the manuscript.

Members of the DEPRESSD PHQ Group contributed:

To data extraction, coding, and synthesis: CH, YW, AK, PMB, ZN, MImran, DBR, KER, MA, AWL. Via the design and conduct of database searches: JTB, LAK. As members of the DEPRESSD Steering Committee, including conception and oversight of collaboration: SG, DM. By contributing included datasets: DA, LA, HRB, ABeraldi, CNB, ABhana, RIB, MHC, JCNC, LFC, DC, AC, FMD, JMdMvG, CDQ, SF, JRWF, DF, ECG, BG, LG, LJG, EPG, BJH, LHantsoo, EEH, MHärter, UH, LHides, SEH, SH, MHudson, TH, MInagaki, HJJ, NJ, MEK, SK, BAK, YK, FL, MAL, HFLA, SIL, ML, SRL, BLöwe, NPL, CL, RAM, BPM, SMS, TNM, KM, JEMN, LN, FLO, PP, AP, SLP, TJQ, ER, SDR, KR, HJR, ISS, MTS, JS, EHS, LSpangenberg, LStafford, SCS, KS, PLLT, MTR, TDT, CMvdFC, TvH, HCvW, LIW, JLW, DW, KW, MY, QZZ, YZ.

All authors, including group authors, provided a critical review and approved the final manuscript. DH is the guarantor; she had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analyses. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding:

This study was funded by the Canadian Institutes of Health Research (CIHR; KRS-134297, PJT-162206). Dr. Levis and Dr. Wu were supported by Fonds de recherche du Québec - Santé (FRQ-S) Postdoctoral Training Fellowships. The primary study by Fischer et al. was funded by the German Federal Ministry of Education and Research (01GY1150). The primary studies by Patten et al. and Prisnie et al. were supported by the Cumming School of Medicine, University of Calgary, and Alberta Health Services through the Calgary Health Trust, as well as the Hotchkiss Brain Institute. Dr. Patten was supported by a Senior Health Scholar award from Alberta Innovates Health Solutions. Dr. Benedetti was supported by a FRQ-S researcher salary award. Dr. Thombs was supported by a Tier 1 Canada Research Chair. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. The primary study by Amtmann et al. was supported by a grant from the Department of Education (NIDRR grant number H133B080025) and by the National Multiple Sclerosis Society (MB 0008). Data collection for the study by Ayalon et al. was supported from a grant from Lundbeck International. The primary study by Khamseh et al. was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary studies by Marrie et al. and Bernstein et al. were supported by CIHR (THC-135234) and Crohn's and Colitis Canada. Dr. Bernstein was supported in part by the Bingham Chair in Gastroenterology. The primary studies by Bhana et al., Kohrt et al. and Nakku et al. were output of the PRogramme for Improving Mental health carE (PRIME) and were supported by the UK Department for International Development (201446). The views expressed do not necessarily reflect the UK Government’s official policies. The primary study by Buji et al. was supported by grants from the UKMMC Fundamental Research Fund (FF-2015-051) and the Fundamental Research Grant Scheme by the Malaysian Ministry of Higher Education (FRGS/2/2014/SKK09/UKM/02/1). Collection of data for the primary study by Zhang et al. was

supported by the European Foundation for Study of Diabetes, the Chinese Diabetes Society, Lilly Foundation, Asia Diabetes Foundation and Liao Wun Yuk Diabetes Memorial Fund. The primary study by Chibanda et al. was supported by a grant from Grand Challenges Canada (0087-04). The primary study by Conway et al. was supported by the Institute of Health and Biomedical Innovation at Queensland University of Technology and the Sigma Theta Tau International Honour Society of Nursing (ID: 8580). The primary study by Grool et al. was supported by a programme grant from The Netherlands Heart Foundation (2007B027). The primary study by Zuithoff et al. was supported by The European Commission (PREDICTQL4-CT2002-00683) and The Netherlands Organization for Scientific Research (ZonMw 016.046.360). The primary study by Martin-Subero et al. was supported in part by a grant from the Spanish Ministry of Health's Health Research Fund (Fondo de Investigaciones Sanitarias, project 97/1184). The primary study by van Heyningen et al. was supported by the Medical Research Council of South Africa (415865), Cordaid Netherlands (Project 103/10002 G Sub 7) and the Truworths Community Foundation Trust, South Africa. The primary study by Fisher et al. was supported by grants from the National Health and Medical Research Council (APP1026550), the Australian Government Department of Social Services - Families, Housing, Community Services and Indigenous Affairs, and the Victorian Department of Education and Early Childhood Development. Dr. Fisher was supported by a Monash Professorial Fellowship and the Jean Hailes Professorial Fellowship, which is supported by a grant to the Jean Hailes Foundation from the H and L Hecht Trust managed by Perpetual Trustees. The primary study by Baron et al. was supported by the National Income Dynamics Study (NIDS), at the University of Cape Town, South Africa. The NIDS is implemented by the Southern Africa Labour and Development Research Unit, and is funded by the Department of Planning, Monitoring and

Evaluation. The funding body was involved in the design of the primary study. Data for the primary study by Gelaye et al. was supported by a grant from the NIH (T37 MD001449). The primary study by Gholizadeh et al. was supported by University of Technology Sydney under UTS Research Reestablishment Grants. The primary study by Green et al. (2018) was supported by a grant from the Duke Global Health Institute (453-0751). Collection of data for the primary study by Hobfoll et al. was made possible in part from grants from NIMH (RO1 MH073687) and the Ohio Board of Regents. Dr. Hall received support from a grant awarded by the Research and Development Administration Office, University of Macau (MYRG2015-00109-FSS). The primary study by Garabiles et al. was supported by the Macao (SAR) Government, through the University of Macau RSKTO grants: MYRG-2014-111. The primary study by Hantsoo et al. was supported by K23 MH107831-02, Brain and Behavior Research Foundation NARSAD Young Investigator Award. The primary study by Haroz et al. was supported by the United States Agency for International Development Victims of Torture Fund: AID-DFD A-00-08-00308. Dr. Haroz was supported by a NIMH T32 predoctoral training grant (MH014592-38) and postdoctoral training grant (MH103210) during the conduct of primary study. Collection of data provided by Drs. Härter and Reuter was supported by the Federal Ministry of Education and Research (grants No. 01 GD 9802/4 and 01 GD 0101) and by the Federation of German Pension Insurance Institute. The primary study by Henkel et al. was funded by the German Ministry of Research and Education. The primary study by Hides et al. was funded by the Perpetual Trustees, Flora and Frank Leith Charitable Trust, Jack Brockhoff Foundation, Grosvenor Settlement, Sunshine Foundation and Danks Trust. Data for the study by Razykov et al. was collected by the Canadian Scleroderma Research Group, which was funded by the CIHR (FRN 83518), the Scleroderma Society of Canada, the Scleroderma Society of Ontario, the

Scleroderma Society of Saskatchewan, Sclérodermie Québec, the Cure Scleroderma Foundation, Inova Diagnostics Inc., Euroimmun, FRQ-S, the Canadian Arthritis Network, and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montréal, QC. Dr. Hudson was supported by a FRQ-S Senior Investigator Award. Collection of data for the primary study by Hyphantis et al. (2014) was supported by grant from the National Strategic Reference Framework, European Union, and the Greek Ministry of Education, Lifelong Learning and Religious Affairs (ARISTEIA-ABREVIATE, 1259). The primary study by Paika et al. was supported by the European Economic Area (EEA) Financial Mechanism 2009–2014 (EEA GR07/3767) and National funds as part of the program "Dissimilarity, Inequality and Social Integration" (132324/I4-25/8/2015). The primary study by Inagaki et al. was supported by the Ministry of Health, Labour and Welfare, Japan. The primary study by Kim et al. was supported by a grant from the Korean Mental Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (HM14C2567), an Institute for Information & Communications Technology Promotion grant funded by the Korea government (MSIP) (B0132- 15-1003: the development of skin adhesive patches for the monitoring and prediction of mental disorders), and by the Original Technology Research Program for Brain Science through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2016M3C7A1947307). Dr. Jetté was supported by a Canada Research Chair in Neurological Health Services Research and an AIHS Population Health Investigator Award. She is the Icahn School of Medicine at Mount Sinai Bludhorn Professor of International Medicine. The primary study by Janssen et al. was supported by the European Regional Development Fund as part of OP-ZUID; the Province of Limburg; the department of Economic Affairs of the Netherlands (31O.041); Stichting the Weijerhorst, the Pearl String Initiative Diabetes; the

Cardiovascular Center Maastricht Cardiovasculair Research Institute Maastricht; School of Nutrition, Toxicology and Metabolism; Stichting Annadal; and Health Foundation Limburg. The primary study by Lamers et al. was funded by the Netherlands Organisation for Health Research and development (grant number 945-03-047). The primary study by Lara et al. was supported by the Consejo Nacional de Ciencia y Tecnología/National Council for Science and Technology (CB-2009-133923-H). The primary study by Liu et al. (2011) was funded by a grant from the National Health Research Institute, Republic of China (NHRI-EX97-9706PI). The primary study by Lotrakul et al. was supported by the Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (grant number 49086). The primary studies by Osório et al. (2012) were funded by Reitoria de Pesquisa da Universidade de São Paulo (grant number 09.1.01689.17.7) and Banco Santander (grant number 10.1.01232.17.9). Dr. Bernd Löwe received research grants from Pfizer, Germany, and from the medical faculty of the University of Heidelberg, Germany (project 121/2000) for the study by Gräfe et al. Dr. Marrie was supported by the Waugh Family Chair in Multiple Sclerosis and the Research Manitoba Chair, and CIHR grants, during the conduct of the study. Dr. Marx was supported by the Department of Defense (W81XWH-08-2- 0100/W81XWH-08-2-0102 and W81XWH-12- 2-0117/W81XWH-12-2-0121). The primary study by Mohd Sidik et al. was funded under the Research University Grant Scheme from Universiti Putra Malaysia, Malaysia and the Postgraduate Research Student Support Accounts of the University of Auckland, New Zealand. The primary study by Santos et al. was funded by the National Program for Centers of Excellence (PRONEX/FAPERGS/CNPq, Brazil). The primary study by Muramatsu et al. (2007) was supported by an educational grant from Pfizer US Pharmaceutical Inc. The primary study by Muramatsu et al. (2018) was supported by grants from Niigata Seiryo University. Dr. Osório was supported by Productivity

Grants (PQ-CNPq-2 -number 301321/2016-7). The primary study by Persoons et al. was partly funded by a grant from the Belgian Ministry of Public Health and Social Affairs and supported by a limited grant from Pfizer Belgium. The primary study by Picardi et al. was supported by funds for current research from the Italian Ministry of Health. The primary study by Wagner et al. was supported by grants U10CA21661, U10CA180868, U10CA180822, and U10CA37422 from the National Cancer Institute. The study was also funded in part by a grant from the Pennsylvania Department of Health. The department specifically declaims responsibility for any analyses, interpretations, or conclusions of the primary study. The primary study by Rancans et al. was supported by The National Research Programme BIOMEDICINE 2014–2017 (5.8.1.). Dr. Shaaban was supported by funding from Universiti Sains Malaysia. The primary study by Shinn et al. was supported by grant NCI K07 CA 093512 and the Lance Armstrong Foundation. The primary study by Spangenberg et al. was supported by a junior research grant from the medical faculty, University of Leipzig. Dr. Stafford received PhD scholarship funding from the University of Melbourne. Dr. Tran was supported by a Monash Bridging Postdoctoral Fellowship. The primary study by Volker et al. was supported by The Netherlands organization for Health Research and Development (ZonMw) and from Achmea SZ, a Dutch insurance company. The study by Wittkampf et al. was funded by The Netherlands Organization for Health Research and Development (ZonMw) Mental Health Program (nos. 100.003.005 and 100.002.021) and the Academic Medical Center/University of Amsterdam. The primary study by Liu et al. (2015) was supported by CIHR (MOP-114970). The primary study by Liu et al. (2016) was supported by Shanghai Municipal Health and Family Planning Commission Bureau-level Project - Preliminary Exploration of Depression Risk Prediction Model for Outpatients in General Hospitals (Project No. 2010105). No other authors reported funding for primary studies

or for their work on this study. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Declaration of Competing Interests:

All authors have completed the ICJME uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years with the following exceptions: Dr. Bernstein declares that he has consulted to Abbvie Canada, Amgen Canada, Bristol Myers Squibb Canada, Roche Canada, Janssen Canada, Pfizer Canada, Sandoz Canada, Takeda Canada, and Mylan Pharmaceuticals. He has also received unrestricted educational grants from Abbvie Canada, Janssen Canada, Pfizer Canada, and Takeda Canada; as well as been on speaker's bureau of Abbvie Canada, Janssen Canada, Takeda Canada and Medtronic Canada, all outside the submitted work. Dr. Chan J CN is a steering committee member and/or consultant of Astra Zeneca, Bayer, Lilly, MSD and Pfizer. She has received sponsorships and honorarium for giving lectures and providing consultancy and her affiliated institution has received research grants from these companies. Dr. Chan LF declares personal fees and non-financial support from Otsuka, Lundbeck, and Johnson and Johnson; and non-financial support from Ortho-McNeil-Janssen, and Menarini, outside the submitted work. Dr. Hegerl declares that within the last three years, he was an advisory board member Janssen and received a research grant from Medice, all outside the submitted work. Dr. Inagaki declares that he has received personal fees from Meiji, Mochida, Takeda, Novartis, Yoshitomi, Pfizer, Eisai, Otsuka, MSD, Sumitomo Dainippon,

Janssen, and Eli Lilly, all outside of the submitted work. Dr. Pugh declares that she received salary support from Pfizer-Astella and Millennium, outside the submitted work. Dr. Rancans declares that he received grants, personal fees and non-financial support from Gedeon Richter; personal fees and non-financial support from Lundbeck, Servier, and Janssen Cilag; personal fees from Zentiva, and Abbvie; outside the submitted work. Dr. Schram declares that the primary study by Janssen et al. was supported by unrestricted grants from Janssen, Novo Nordisk, and Sanofi. Dr. Wagner declares that she receives personal fees from Celgene, outside the submitted work. All authors declare no other relationships or activities that could appear to have influenced the submitted work. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

References

[1] C. Goetz *et al.*, "Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales," *Journal of Clinical Epidemiology*, vol. 66, no. 7. pp. 710–718, 2013. doi: 10.1016/j.jclinepi.2012.12.015.

[2] J. Coste, F. Guillemin, J. Pouchot, and J. Fermanian, "Methodological approaches to shortening composite measurement scales," *J. Clin. Epidemiol.*, vol. 50, no. 3, pp. 247–252, 1997, doi: 10.1016/S0895-4356(96)00363-0.

[3] P. M. Kruyen, W. H. M. Emons, and K. Sijtsma, "On the Shortcomings of Shortened Tests: A Literature Review," *International Journal of Testing*, vol. 13, no. 3. pp. 223–248, 2013. doi: 10.1080/15305058.2012.703734.

[4] P. C. Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, "Issues and strategies for reducing the length of self-report scales," *Pers. Psychol.*, vol. 55, no. 1, pp. 167–194, 2002.

[5] K. Kroenke and R. L. Spitzer, *The PHQ-9: a new depression diagnostic and severity measure*. SLACK Incorporated Thorofare, NJ, 2002.

[6] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disord.*, vol. 114, no. 1–3, pp. 163–173, 2009.

[7] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001.

[8] B. Levis, A. Benedetti, and B. D. Thombs, "Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis," *bmj*, vol. 365, 2019.

[9] M. Ishihara *et al.*, "Shortening self-report mental health symptom measures through optimal test assembly methods: Development and validation of the Patient Health Questionnaire-Depression-4," *Depress. Anxiety*, vol. 36, no. 1, pp. 82–92, 2019.

[10] W. J. Van der Linden, *Linear models for optimal test design*. Springer Science & Business Media, 2006.

[11] J.-T. Kuhn and T. Kiefer, "Optimal test assembly in practice," *Z. Für Psychol.*, 2015.

[12] D. Harel *et al.*, "Shortening patient-reported outcome measures through optimal test assembly: application to the social appearance anxiety scale in the scleroderma patient-centered intervention network cohort," *BMJ Open*, vol. 9, no. 2, p. e024010, 2019.

[13] D. Harel *et al.*, "Shortening the Edinburgh Postnatal Depression Scale using Optimal Test Assembly Methods: Development of the EPDS-Dep-5," *Acta Psychiatr. Scand.*, 2020.

[14] D. Harel and M. Baron, "Methods for shortening patient-reported outcome measures," *Stat. Methods Med. Res.*, vol. 28, no. 10–11, pp. 2992–3011, 2019.

[15] A. W. Levis *et al.*, "Using optimal test assembly methods for shortening patient-reported outcome measures: Development and Validation of the Cochin Hand Function Scale-6: A scleroderma patient-centered intervention network cohort study," *Arthritis Care Res.*, vol. 68, no. 11, pp. 1704–1713, 2016.

[16] S. Li *et al.*, "Nonrestorative sleep scale: a reliable and valid short form of the traditional Chinese version," *Qual. Life Res.*, vol. 29, no. 9, pp. 2585–2592, 2020.

[17] S. Li *et al.*, "A Short Form of the Chinese Version of the Weinstein Noise Sensitivity Scale through Optimal Test Assembly," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 3, p. 879, 2021.

[18] Z. Negeri *et al.*, "Accuracy of the Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: an updated systematic review and individual participant data meta-analysis," Under Review.
[19] B. D. Thombs *et al.*, "The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses," *Syst. Rev.*, vol. 3, no. 1, pp. 1–16, 2014.
[20] A. P. Association, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.
[21] A. P. American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (DSM-IV)*, vol. 886. Washington, DC: American psychiatric association Washington, 1994.
[22] *Diagnostic and statistical manual of mental disorders: DSM-III*, 3rd, revised ed. Washington, DC: American Psychiatric Association, 1987.
[23] *Diagnostic and statistical manual of mental disorders: DSM-III*, 4th, revised ed. Washington, DC: American Psychiatric Association, 2000.
[24] W. H. Organization, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization, 1992.
[25] J. McGowan, M. Sampson, D. M. Salzwedel, E. Cogo, V. Foerster, and C. Lefebvre, "PRESS peer review of electronic search strategies: 2015 guideline statement," *J. Clin. Epidemiol.*, vol. 75, pp. 40–46, 2016.
[26] R. D. Riley, S. R. Dodd, J. V. Craig, J. R. Thompson, and P. R. Williamson, "Meta-analysis of diagnostic test studies using individual patient data and aggregate data," *Stat. Med.*, vol. 27, no. 29, pp. 6111–6136, 2008.
[27] B. Levis *et al.*, "Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews," *Br. J. Psychiatry*, vol. 212, no. 6, pp. 377–385, 2018.
[28] B. Levis *et al.*, "Comparison of major depression diagnostic classification probability using the SCID, CIDI, and MINI diagnostic interviews among women in pregnancy or postpartum: An individual participant data meta-analysis," *Int. J. Methods Psychiatr. Res.*, vol. 28, no. 4, p. e1803, 2019.
[29] Y. Wu *et al.*, "Probability of major depression diagnostic classification based on the SCID, CIDI and MINI diagnostic interviews controlling for Hospital Anxiety and Depression Scale–Depression subscale scores: an individual participant data meta-analysis of 73 primary studies," *J. Psychosom. Res.*, vol. 129, p. 109892, 2020.
[30] Y. Wu, B. Levis, J. P. Ioannidis, A. Benedetti, and B. D. Thombs, "Probability of Major Depression Classification Based on the SCID, CIDI, and MINI Diagnostic Interviews: A Synthesis of Three Individual Participant Data Meta-Analyses," *Psychother. Psychosom.*, vol. 90, no. 1, pp. 28–40, 2021.
[31] Y. Wu *et al.*, "Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis," *Psychol. Med.*, vol. 50, no. 8, pp. 1368–1380, 2020.
[32] R. Van der Leeden, F. Busing, and E. Meijer, "Bootstrap Methods for Two-Level Models: Technical Report PRM 97-04," *Leiden Univ. Dep. Psychol. Leiden Neth.*, 1997.
[33] R. Van der Leeden, E. Meijer, and F. M. Busing, "Resampling multilevel models," in *Handbook of multilevel analysis*, Springer, 2008, pp. 401–433.

[34] E. Walker and A. S. Nowacki, “Understanding equivalence and noninferiority testing,” *J. Gen. Intern. Med.*, vol. 26, no. 2, pp. 192–196, 2011.
[35] R. C. Team, “R: A language and environment for statistical computing,” 2013.
[36] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *ArXiv Prepr. ArXiv14065823*, 2014.
[37] E. Arthurs, R. J. Steele, M. Hudson, M. Baron, B. D. Thombs, and (CSRG) Canadian Scleroderma Research Group, “Are scores on English and French versions of the PHQ-9 comparable? An assessment of differential item functioning,” *PloS One*, vol. 7, no. 12, p. e52028, 2012.
[38] A. Teymoori *et al.*, “Measurement invariance of assessments of depression (PHQ-9) and anxiety (GAD-7) across sex, strata and linguistic backgrounds in a European-wide sample of patients after Traumatic Brain Injury,” *J. Affect. Disord.*, vol. 262, pp. 278–285, 2020.
[39] H. Reich, W. Rief, and E. Brahler, “Cross-cultural validation of the German and Turkish versions of the PHQ-9: an IRT approach. BMC Psychol. 2018; 6 (26),” Epub 2018/06/07. PubMed PMID: 29871664.

FIGURES

Figure 1: Flow diagram of study selection process

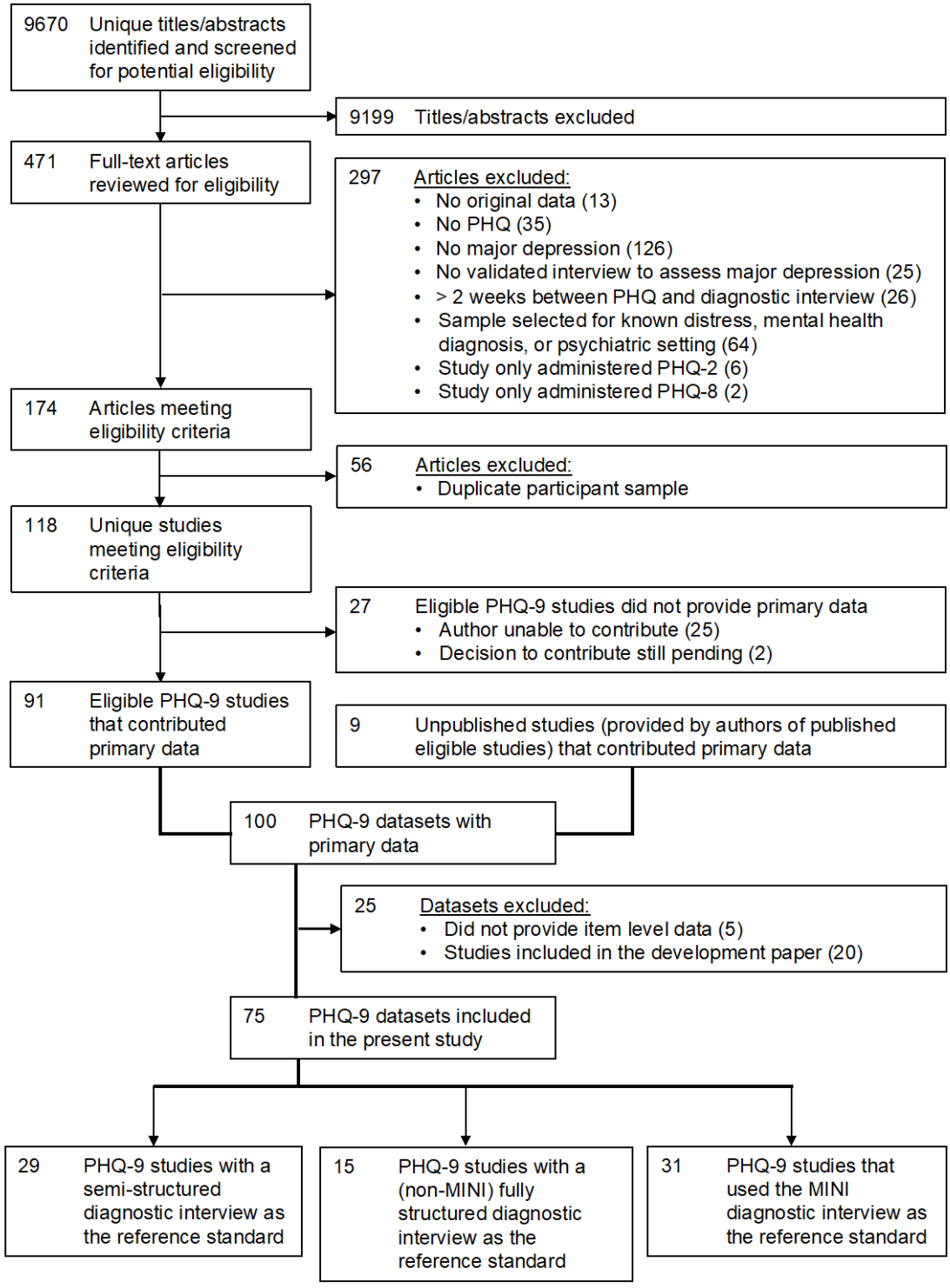
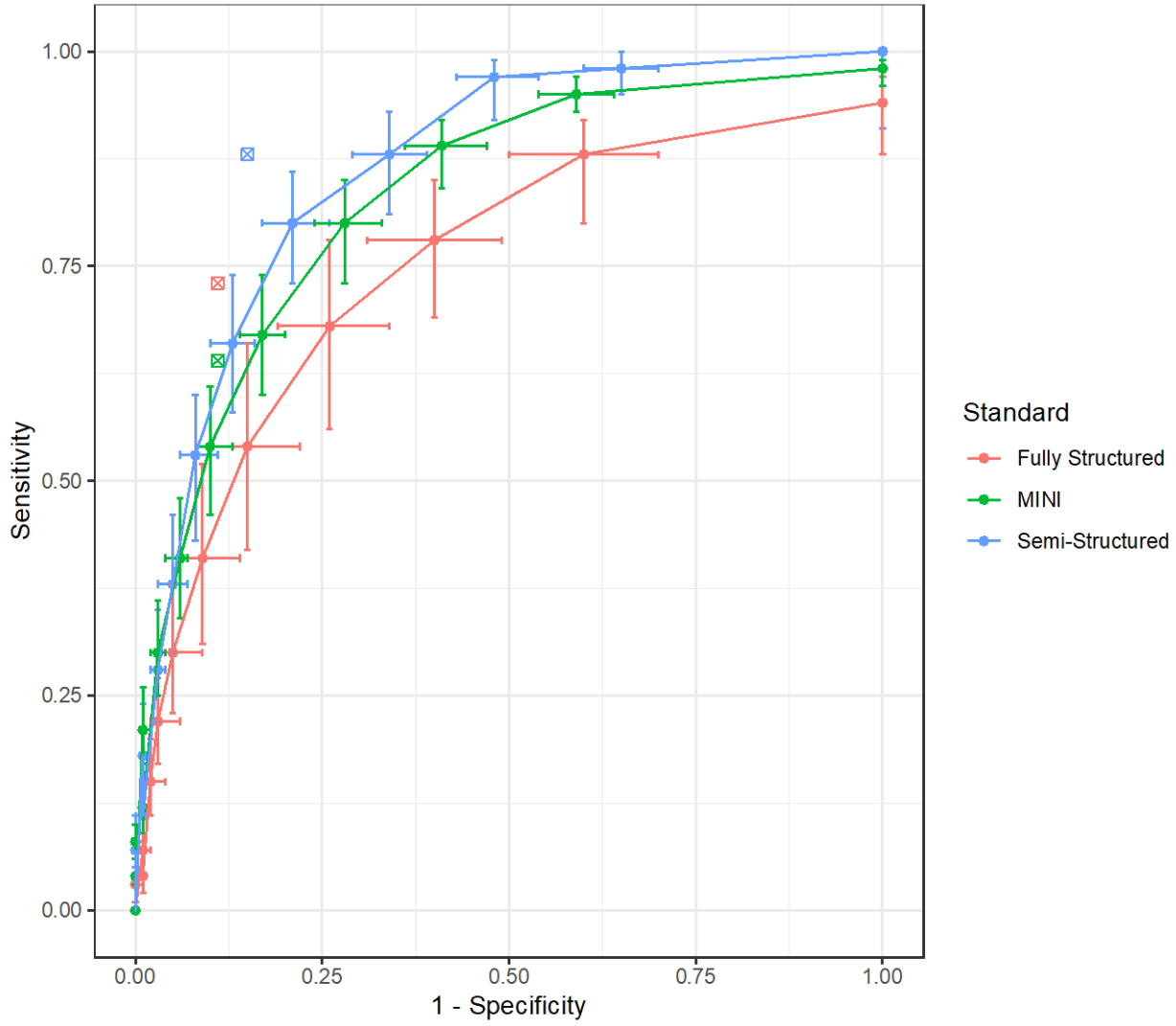


Figure 2: Receiver-operating curve for each reference standard category. Points represent cutoffs of 0 (right) to 12 (left) for each reference standard category. X marks the PHQ-9 cutoff of ≥ 10.



TABLES

Table 1: Demographics of the study sample for patients with and without major depression

Sociodemographic variables	Total (N=34,698)	Participants with Major Depression (N=3,392)	Participants without Major Depression (N=31,306)
Age in years, *mean [median] ± SD (range)*[1]	47.7 [48] ± 16.3 (18, 98)	46.4 [45] ± 16.3 (18, 94)	48.9 [48] ± 16.3 (18, 98)
Women, *n (%)*[2]	20678	2351 (11.4)	18327 (88.6)
Men, *n (%)*[2]	13998	1038 (7.4)	12960 (92.6)
PHQ-9 score, *mean [median] ± SD (range)*	4.9 [3] ± 5.2 (0, 27)	13.1 [13] ± 6.3 (0, 27)	4.0 [3] ± 4.2 (0, 27)
Country, *n (%)*			
Netherlands	7049	494 (7.0)	6555 (93.0)
Canada	5215	190 (3.6)	5025 (96.4)
South Korea	3071	205 (6.7)	2866 (93.3)
South Africa	2300	299 (13.0)	2001 (87.0)
China	2096	136 (6.5)	1960 (93.5)
Germany	1605	147 (9.2)	1458 (90.8)
Taiwan	1532	50 (3.3)	1482 (96.7)
Latvia	1467	147 (10.0)	1320 (90.0)
USA	1247	166 (13.3)	1081 (86.7)
Greece	1036	262 (25.3)	774 (74.7)
Spain	1003	83 (8.3)	920 (91.7)
Other[3]	7077	1213 (17.1)	5864 (82.9)
Language, *n (%)*[4]			
English	8073	562 (7.0)	7511 (93.0)
Dutch	7222	522 (7.2)	6700 (92.8)
Chinese	3597	164 (4.6)	3433 (95.4)
Korean	3071	205 (6.7)	2866 (93.3)
South African languages	1838	211 (11.5)	1627 (88.5)
German	1605	147 (9.2)	1458 (90.8)
Spanish	1540	181 (11.8)	1359 (88.2)
Greek	1036	262 (25.3)	774 (74.7)
Other[5]	6611	1130 (17.1)	5481 (82.9)
General Care Setting, *n (%)*			
Outpatient care	17624	2250 (12.8)	15374 (87.2)
Inpatient care	2781	331 (11.9)	2450 (88.1)
Non-medical setting	14163	806 (5.7)	13357 (94.3)
Outpatient/inpatient mixed sample	130	5 (3.8)	125 (96.2)
Diagnostic Interview, *n (%)*			
SCID	6187	873 (14.1)	5314 (85.9)
CIDI	11810	860 (7.3)	10950 (92.7)
SCAN	1532	50 (3.3)	1482 (96.7)
MINI	14870	1596 (10.7)	13274 (89.3)
CIS-R	299	13 (4.3)	286 (95.7)
Classification system, *n (%)*			
ICD-10	909	86 (9.5)	823 (90.5)

DSM-III	1107	104 (9.4)	1003 (90.6)
DSM-IV	31771	3089 (9.7)	28682 (90.3)
DSM-V	911	113 (12.4)	798 (87.6)

[1]N missing = 31 participants with major depression, 216 participants without major depression
[2]N missing = 3 participants with major depression, 19 participants without major depression
[3]Other countries: Ethiopia, Japan, Australia, Brazil, Singapore, Malaysia, India, Israel, Mexico, Thailand, Zimbabwe, Argentina, Uganda, Iran, Kenya, Belgium, Italy, UK, Myanmar, Nepal, Hong Kong China.
[4]N missing = 8 for MDD, 97 for non-MDD
[5]Other Languages: Amharic, Latvian, Japanese, Russian, Portuguese, Malay, Indian languages (unspecified), Malay or English, Thai, Shona, Hebrew, Farsi, Kiswahili, Italian, Burmese, Nepali, Malay, Chinese or Tamil, Filipino, Arabic, French

Table 2: Sensitivity and specificity for each PHQ-Dep-4 cutoff and the PHQ-9 cutoff of ≥ 10

	SEMI-STRUCTURED REFERENCE STANDARD: N studies = 29, N participants = 7719, N major depression = 923				**FULLY STRUCTURED REFERENCE STANDARD: N studies = 15, N participants = 12,109, N major depression = 873**				**MINI[2] REFERENCE STANDARD: N studies = 31, N participants = 14,870, N major depression = 1596**			
Cutoff PHQ-Dep-4	**sensitivity**	**95% CI**	**specificity**	**95% CI**	**sensitivity**	**95% CI**	**specificity**	**95% CI**	**sensitivity**	**95% CI**	**specificity**	**95% CI**
>= 1	1.00	(0.91, 1.00)	0.35	(0.30, 0.40)	0.94	(0.88, 0.97)	0.40	(0.30, 0.50)	0.98	(0.96, 0.99)	0.41	(0.36, 0.46)
>= 2	0.98	(0.95, 1.00)	0.52	(0.46, 0.57)	0.88	(0.80, 0.92)	0.60	(0.51, 0.69)	0.95	(0.93, 0.97)	0.59	(0.53, 0.64)
>= 3	0.97	(0.92, 0.99)	0.66	(0.61, 0.71)	0.78	(0.69, 0.85)	0.74	(0.66, 0.81)	0.89	(0.84, 0.92)	0.72	(0.67, 0.76)
>= 4	0.88	(0.81, 0.93)	0.79	(0.74, 0.83)	0.68	(0.56, 0.78)	0.85	(0.78, 0.90)	0.80	(0.73, 0.85)	0.83	(0.80, 0.86)
>= 5	0.80	(0.73, 0.86)	0.87	(0.84, 0.90)	0.54	(0.42, 0.66)	0.91	(0.86, 0.94)	0.67	(0.60, 0.74)	0.90	(0.87, 0.92)
>= 6	0.66	(0.58, 0.74)	0.92	(0.89, 0.94)	0.41	(0.31, 0.52)	0.95	(0.91, 0.97)	0.54	(0.46, 0.61)	0.94	(0.93, 0.96)
>= 7	0.52	(0.43, 0.60)	0.95	(0.93, 0.97)	0.30	(0.23, 0.38)	0.97	(0.94, 0.98)	0.41	(0.34, 0.48)	0.97	(0.96, 0.98)
>= 8[1]	0.38	(0.30, 0.46)	0.97	(0.96, 0.98)	0.22	(0.17, 0.27)	0.98	(0.96, 0.99)	0.30	(0.25, 0.36)	0.99	(0.98, 0.99)
>= 9	0.28	(0.22, 0.35)	0.99	(0.98, 0.99)	0.15	(0.11, 0.20)	0.99	(0.98, 0.99)	0.21	(0.17, 0.26)	0.99	(0.99, 0.99)
>= 10	0.18	(0.13, 0.24)	0.99	(0.99, 1.00)	0.07	(0.04, 0.12)	0.99	(0.99, 1.00)	0.12	(0.09, 0.16)	1.00	(0.99, 1.00)
>= 11	0.11	(0.08, 0.16)	1.00	(0.99, 1.00)	0.04	(0.02, 0.07)	1.00	(0.99, 1.00)	0.08	(0.06, 0.10)	1.00	(1.00, 1.00)
>= 12	0.07	(0.05, 0.11)	1.00	(1.00, 1.00)	0.03	(0.01, 0.06)	1.00	(1.00, 1.00)	0.04	(0.03, 0.06)	1.00	(1.00, 1.00)
PHQ-9 >= 10	0.88	(0.81, 0.93)	0.85	(0.80, 0.88)	0.64	(0.50, 0.76)	0.89	(0.83, 0.93)	0.73	(0.66, 0.79)	0.89	(0.86, 0.91)

[1]BOBYQA optimizer was used to ensure model convergence for the semi-structured reference category, as the model with the default optimizer did not converge

[2]MINI: Mini International Neuropsychiatric Interview

Table 3: Results of the equivalence tests between the accuracy of the PHQ-Dep-4 and PHQ-9 ≥ 10

All studies (N studies = 75, N participants = 34,698, N major depression = 3392)				
Cutoff	**Sensitivity Difference** (PHQ-Dep-4 - PHQ-9 >=10)	**95% CI**	**Specificity Difference** (PHQ-Dep-4 - PHQ-9 >=10)	**95% CI**
PHQ-Dep-4 >= 1	0.21	(0.14, 0.25)	-0.49	(-0.52, -0.46)
PHQ-Dep-4 >= 2	0.18	(0.13, 0.22)	-0.31	(-0.34, -0.28)
PHQ-Dep-4 >= 3	0.13	(0.09, 0.16)	-0.17	(-0.19, -0.15)
PHQ-Dep-4 >= 4	0.03	(0.00, 0.06)	-0.05	(-0.07, -0.04)
PHQ-Dep-4 >= 5	-0.07	(-0.11, -0.05)	0.02	(0.01, 0.03)
PHQ-Dep-4 >= 6	-0.22	(-0.27, -0.19)	0.06	(0.05, 0.08)
PHQ-Dep-4 >= 7	-0.35	(-0.41, -0.33)	0.09	(0.08, 0.11)
PHQ-Dep-4 >= 8	-0.47	(-0.53, -0.45)	0.11	(0.09, 0.13)
PHQ-Dep-4 >= 9	-0.55	(-0.62, -0.53)	0.12	(0.10, 0.14)
PHQ-Dep-4 >= 10	-0.65	(-0.72, -0.62)	0.12	(0.10, 0.15)
PHQ-Dep-4 >= 11	-0.70	(-0.77, -0.67)	0.13	(0.10, 0.15)
PHQ-Dep-4 >= 12	-0.73	(-0.80, -0.69)	0.13	(0.10, 0.15)

SUPPLEMENTARY MATERIALS

eMethods1: 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=2000-20181.

Supplementary Table 1a. Characteristics of included primary studies (N=75)

First Author, Year	Country	Recruited Population	Diagnostic Interview	Classification System	Total N	Major Depression N (%)
Semi-structured Interviews						
Amtmann, 2015[1]	USA	Multiple sclerosis patients	SCID	DSM-IV	164	48 (29)
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)
Bernstein, 2018[4]	Canada	IBD patients	SCID	DSM-IV	240	21 (9)
Bhana, 2015[5]	South Africa	Chronic care patients	SCID	DSM-IV	679	78 (11)
Chagas, 2013[6]	Brazil	Outpatients with Parkinson's Disease	SCID	DSM-IV	84	19 (23)
Chibanda, 2016[7]	Zimbabwe	A primary care population with high HIV prevalence	SCID	DSM-IV	264	149 (56)
Fischer, 2014[8]	Germany	Heart failure patients	SCID	DSM-IV	194	11 (6)
Gräfe, 2004[9]	Germany	Medical and psychosomatic outpatients	SCID	DSM-IV	494	67 (14)
Green, 2017[10]	USA	Returning veterans	SCID	DSM-V	176	22 (13)
Green, 2018[11]	Kenya	Pregnant women and new mothers	SCID	DSM-V	192	10 (5)
Haroz, 2017[12]	Myanmar	Primary care patients	SCID	DSM-IV	132	29 (22)
Hitchon, 2019[13a]	Canada	Rheumatoid arthritis patients	SCID	DSM-IV	148	16 (11)
Khamseh, 2011[14]	Iran	Type 2 diabetes patients	SCID	DSM-IV	122	47 (39)
Kwan, 2012[15]	Singapore	Post-stroke inpatients undergoing rehabilitation	SCID	DSM-IV-TR	113	3 (3)
Lara, 2015[16]	Mexico	Pregnant women during the third trimester of pregnancy	SCID	DSM-IV	280	29 (10)
Liu, 2011[17]	Taiwan	Primary care patients	SCAN	DSM-IV	1532	50 (3)
Marrie, 2018[18]	Canada	Multiple sclerosis patients	SCID	DSM-IV	244	25 (10)

Martin-Subero, 2017[19]	Spain	Medical inpatients	SCID	DSM-III	1003	83 (8)
Osório, 2009[20]	Brazil	Women in primary care	SCID	DSM-IV	177	60 (34)
Osório, 2012[21]	Brazil	Inpatients from various clinical wards	SCID	DSM-IV	86	28 (33)
Patten, 2015[22]	Canada	Multiple sclerosis patients	SCID	DSM-IV	143	20 (14)
Picardi, 2005[23]	Italy	Inpatients with skin diseases	SCID	DSM-IV	138	12 (9)
Prisnie, 2016[24]	Canada	Stroke and transient ischemic attack patients	SCID	DSM-IV	114	11 (10)
Quinn, Unpublished[a]	UK	Stroke patients	SCID	DSM-V	135	15 (11)
Shinn, 2017[25]	USA	Cancer patients	SCID	DSM-IV	124	5 (4)
Spangenberg, 2015[26]	Germany	Primary care patients	SCID	DSM-IV	160	1 (1)
Wagner, 2017[27]	USA	Patients starting radiotherapy for the first diagnosis of any tumor	SCID	DSM-IV	54	6 (11)
Wittkampf, 2009[28]	The Netherlands	Primary care patients at risk for depression	SCID	DSM-IV	260	45 (17)
Fully Structured Interviews						
Azah, 2005[29]	Malaysia	Adults attending family medicine clinics	CIDI	ICD-10	180	30 (17)
de Man-van Ginkel, 2012[30]	The Netherlands	Stroke patients	CIDI	DSM-IV	382	54 (14)
Fisher, 2016[31]	Australia	Primiparous women less than 6 weeks postpartum	CIDI	DSM-IV	357	4 (1)
Gelaye, 2014[32]	Ethiopia	Outpatients at a general hospital	CIDI	DSM-IV	923	162 (18)
Grool, 2011[33]	The Netherlands	Non-demented patients with symptomatic atherosclerotic disease	CIDI	DSM-IV	477	22 (5)
Hahn, 2006[34]	Germany	Patients with chronic illnesses from rehabilitation centers	CIDI	DSM-IV	211	18 (9)
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	144	42 (29)
Kim, 2017[37]	South Korea	Randomly selected adults	CIDI	DSM-IV	3071	205 (7)
Kohrt, 2016[38]	Nepal	Primary care patients	CIDI	DSM-IV	125	17 (14)

Liu, 2015[39]	Canada	Working population	CIDI	DSM-IV	4182	91 (2)
Mohd Sidik, 2012[40]	Malaysia	Primary care patients	CIDI	DSM-IV	146	31 (21)
Patel, 2008[41]	India	Primary care patients	CIS-R	ICD-10	299	13 (4)
Razykov, 2013[42]	Canada	Patients with systemic sclerosis	CIDI	DSM-IV	144	6 (4)
Zuithoff, 2009[43]	The Netherlands	General practice patients	CIDI	DSM-IV	1038	135 (13)
Mini International Neuropsychiatric Interviews (MINI)						
Akena, 2013[44]	Uganda	HIV/AIDS patients	MINI	DSM-IV	91	11 (12)
Baron, 2017[45]	South Africa	Xhosa, Afrikaans and Zulu-speaking general population	MINI	DSM-IV	851	93 (11)
Buji, 2018[46]	Malaysia	Patients with systemic lupus erythematosus	MINI	DSM-IV	130	5 (4)
Cholera, 2014[47]	South Africa	Patients undergoing routine HIV counseling and testing at a primary health care clinic	MINI	DSM-IV	397	47 (12)
Conway, 2016[48]	Australia	Heart transplant recipients	MINI	DSM-IV	26	2 (8)
de la Torre, 2016[49]	Argentina	Hospitalized general medical patients	MINI	DSM-IV	257	69 (27)
Garabiles, Unpublished[a]	China	Female Filipino domestic workers in Macao	MINI	DSM-IV	99	39 (39)
Gholizadeh, 2019[50a]	Iran	Coronary artery disease patients	MINI	DSM-IV	79	12 (15)
Hantsoo, 2017[51]	USA	General population	MINI	DSM-IV	321	19 (6)
Hides, 2007[52]	Australia	Injection drug users accessing a needle and syringe program	MINI	DSM-IV	103	47 (46)
Hyphantis, 2011[53]	Greece	Patients with various rheumatologic disorders	MINI	DSM-IV	213	69 (32)
Hyphantis, 2014[54]	Greece	Patients with chronic illnesses presenting at the emergency department	MINI	DSM-IV	349	95 (27)
Inagaki, 2013[55]	Japan	Internal medicine outpatients	MINI	DSM-III-R	104	21 (20)
Janssen, 2016[56]	The Netherlands	General population and Type 2 diabetes patients	MINI	DSM-IV	4695	156 (3)

Lamers, 2008[57]	The Netherlands	Elderly primary care patients with diabetes mellitus or chronic obstructive pulmonary disease	MINI	DSM-IV	104	59 (57)
Levin-Aspenson, 2017[58]	USA	General population	MINI	DSM-V	408	66 (16)
Liu, 2016[59]	China	Primary care patients	MINI	DSM-IV	1997	97 (5)
Lotrakul, 2008[60]	Thailand	Outpatients	MINI	DSM-IV	278	19 (7)
Muramatsu, 2007[61]	Japan	Primary care patients	MINI	DSM-IV	116	32 (28)
Muramatsu, 2018[62]	Japan	Primary care patients	MINI	DSM-IV	152	46 (30)
Nakku, 2016[63]	Uganda	Primary patients and hospital outpatients	MINI	DSM-IV	153	84 (55)
Paika, 2017[64]	Greece	Patients with long term medical conditions	MINI	DSM-IV	474	98 (21)
Persoons, 2001[65]	Belgium	Inpatients and patients at gastroenterological and hepatology wards	MINI	DSM-IV	173	28 (16)
Rancans, 2018[66]	Latvia	Primary care patients	MINI	DSM-IV	1467	147 (10)
Santos, 2013[67]	Brazil	General population	MINI	DSM-IV	196	25 (13)
Stafford, 2007[68]	Australia	Inpatients with coronary artery disease who had undergone surgery	MINI	DSM-IV	193	35 (18)
Sung, 2013[69]	Singapore	Primary care patients	MINI	DSM-IV	399	12 (3)
Suzuki, 2015[70]	Japan	Outpatients in general medicine department	MINI	DSM-IV	511	42 (8)
van Heyningen, 2018[71]	South Africa	Pregnant women	MINI	DSM-IV	373	81 (22)
Volker, 2016[72]	The Netherlands	Employees on sickness leave	MINI	DSM-IV	93	23 (25)
Zhang, 2013[73]	Hong Kong, China	Type 2 diabetes patients	MINI	DSM-IV	68	17 (25)

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 Neuropsychiatric Diagnostic Interview; 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 1b. Characteristics of eligible primary studies not included in the present study (N=32)

First Author, Year	Country	Recruited Population	Diagnostic Interview	Classification System	Total N	Major Depression N (%)
Semi-structured Interviews						
Alamri, 2017[74a]	Saudi Arabia	Hospitalized elderly in medical and surgical wards	SCID	DSM-IV	199	24 (12)
Bailer, 2016[75]	Germany	Healthy participants and cognitive behaviour therapy outpatients	SCID	DSM-IV	200	68 (34)
Becker, 2002[76]	Saudi Arabia	Primary care patients	SCID	DSM-III-R	173	NR[a]
Brodey, 2016[77]	USA	Perinatal women	SCID	DSM-IV	879	NR[a]
Chen, 2013[78]	China	Primary care populations	SCID	DSM-IV	280	NR[a]
Chen, 2012[79]	China	Adults over 60 in primary care	SCID	DSM-IV	262	97 (37)
Fann, 2005[80a]	USA	Inpatients with traumatic brain injury	SCID	DSM-IV	135	45 (33)
Irmak, 2017[81]	Turkey	Battered women	SCID	DSM-V	150	63 (42)
Lai, 2010[82]	China	Men with postpartum wives	SCID	DSM-IV	551	8 (1)
Limon, 2016[83]	USA	Latino farmworkers	SCID	DSM-IV	99	NR[a]
Liu, 2016[84]	China	Rural elderly population	SCID	DSM-IV	839	57 (7)
Nacak, 2017[85]	Germany	Patients with somatoform pain disorder	SCID	DSM-IV	130	36 (28)
Navinés, 2012[86]	Spain	Chronic hepatitis C patients	SCID	DSM-IV	500	32 (6)
Phelan, 2010[87]	USA	Elderly primary care patients	SCID	DSM-IV	69	8 (12)
Thompson, 2011[88]	USA	Parkinson's patients	SCID	DSM-IV	214	30 (14)
Vöhringer, 2013[89a]	Chile	Primary care patients	SCID	DSM-IV	190	59 (31)
Watnick, 2005[90]	USA	Long term dialysis patients	SCID	DSM-IV	62	12 (19)
Fully Structured Interviews						

Al-Ghafri, 2014[91]	Oman	Medical trainees	CIDI	NR	131	NR[a]
Haddad, 2013[92]	UK	Coronary heart disease patients	CIS-R	ICD-10	730	32 (4)
Ikin, 2016[93]	Australia	Veterans of the Gulf War	CIDI	DSM-IV	1356	NR[a]
Valencia-Garcia, 2017[94]	USA	Mexican American women	CIDI	DSM-IV	205	40 (20)
Wang, 2015[95]	China	Cardiovascular outpatients	CIDI	DSM-IV	201	42 (21)
Mini International Neuropsychiatric Interviews (MINI)						
Choi, 2015[96]	Canada	HIV patients	MINI	DSM-IV	190	29 (15)
Griffith, 2015[97]	USA	Patients with epilepsy	MINI	DSM-IV and ICD-10	114	20 (18)
Persoons, 2003[98]	Belgium	Otorhinolaryngology outpatients	MINI	DSM-IV	97	16 (16)
Rathore, 2014[99]	USA	Patients with epilepsy	MINI	DSM-IV	158	36 (23)
Scott, 2011[100]	USA	Chronic hepatitis C patients	MINI	DSM-IV and ICD-10	30	NR[a]
Seo, 2015[101]	South Korea	Migrane patients	MINI	DSM-IV	132	39 (30)
van Steenbergen-Weijenburg, 2010[102a]	The Netherlands	Diabetes patients	MINI	DSM-IV	196	37 (19)
Wang, 2014[103a]	China	General population	MINI	DSM-IV	1036	28 (3)
Woldetensay, 2018[104]	Ethiopia	Pregnant women	MINI	DSM-IV	216	28 (13)
Xiong, 2014[105]	China	Outpatients with multiple somatic symptoms	MINI	DSM-IV	398	116 (29)

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; NR: Not Reported; SCID: Structured Clinical Interview for DSM Disorders; UK: United Kingdom; USA: United States of America.

[a]Studies contributed data but were excluded for not having item scores.

Supplementary Table 1c. Characteristics of eligible primary studies included in the PHQ-dep-4 development paper (N=20)

First Author, Year	Country	Recruited Population	Diagnostic Interview	Classification System	Total N	Major Depression N (%)
Semi-structured Interviews						
Amoozegar, 2017[106]	Canada	Migraine patients	SCID	DSM-IV	203	49 (24)
Bombardier, 2012[107]	USA	Inpatients with spinal cord injuries	SCID	DSM-IV	160	14 (9)
Eack, 2006[108]	USA	Women seeking psychiatric services for their children at two mental health centers	SCID	DSM-IV	48	12 (25)
Fiest, 2014[109]	Canada	Epilepsy outpatients	SCID	DSM-IV	169	23 (14)
Gjerdingen, 2009[110]	USA	Mothers registering their newborns for well-child visits at medical or pediatric clinics	SCID	DSM-IV	419	19 (5)
Lambert, 2015[111]	Australia	Cancer patients	SCID	DSM-IV	147	21 (14)
McGuire, 2013[112]	USA	Acute coronary syndrome inpatients	DISH	DSM-IV	100	9 (9)
Richardson, 2010[113]	USA	Older adults undergoing in-home aging services care management assessment	SCID	DSM-IV	377	95 (25)
Rooney, 2013[114]	UK	Patients with cerebral glioma	SCID	DSM-IV	126	14 (11)
Sidebottom, 2012[115]	USA	Pregnant women	SCID	DSM-IV	246	12 (5)
Simning, 2012[116]	USA	Older adults living in public housing	SCID	DSM-IV	190	10 (5)
Turner, 2012[117]	Australia	Stroke patients	SCID	DSM-IV	72	13 (18)
Turner, Unpublished[a]	Australia	Cardiac rehabilitation patients	SCID	DSM-IV	51	4 (8)
Twist, 2013[118]	UK	Type 2 diabetes outpatients	SCAN	DSM-IV	360	80 (22)
Williams, 2012[119]	USA	Parkinson's Disease patients	SCID	DSM-IV	235	61 (26)
Fully Structured Interviews						

Arroll, 2010[120]	New Zealand	Primary care patients	CIDI	DSM-IV	2528	156 (6)
Delgadillo, 2011[121]	UK	Injecting drug users	CIS-R	ICD-10	103	51 (50)
Kiely, 2014[122]	Australia	Community sample of adults	CIDI	ICD-10	822	33 (4)
Pence, 2012[123]	Cameroon	HIV-infected patients	CIDI	DSM-IV	398	11 (3)
Thombs, 2008[124]	USA	Outpatients with coronary artery disease	C-DIS	DSM-IV	1006	221 (22)

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 Neuropsychiatric Diagnostic Interview; 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 MATERIAL REFERENCES

1. Amtmann D, Bamer AM, Johnson KL, et al. A comparison of multiple patient reported outcome measures in identifying major depressive disorder in people with multiple sclerosis. J Psychosom Res. 2015;79:550-557.
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, et al. 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. Bernstein CN, Zhang L, Lix LM, et al. The validity and reliability of screening measures for depression and anxiety disorders in inflammatory bowel disease. Inflamm Bowel Dis. 2018;24:1867-1875.
5. Bhana A, Rathod SD, Selohilwe O, et al. The validity of the Patient Health Questionnaire for screening depression in chronic care patients in primary health care in South Africa. BMC psychiatry. 2015;15:118.
6. 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;42:645-49.
7. Chibanda D, Verhey R, Gibson LJ, et al. Validation of screening tools for depression and anxiety disorders in a primary care population with high HIV prevalence in Zimbabwe. J Affect Disord. 2016;198:50-55.
8. 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.
9. Gräfe K, Zipfel S, Herzog W, et al. Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. Diagnostica. 2004;50:171-181.
10. Green JD, Annunziata A, Kleiman SE, et al. Examining the diagnostic utility of the DSM-5 PTSD symptoms among male and female returning veterans. Depress Anxiety. 2017;34:752-760.
11. Green EP, Tuli H, Kwobah E, et al. Developing and validating a perinatal depression screening tool in Kenya blending Western criteria with local idioms: A mixed methods study. J Affect Disord. 2018;228:49-59.
12. Haroz EE, Bass J, Lee C, et al. Development and cross-cultural testing of the International Depression Symptom Scale (IDSS): a measurement instrument designed to represent global presentations of depression. Glob Ment Health. 2017;4.
13. Hitchon CA, Zhang L, Peschken CA, et al. The validity and reliability of screening measures for depression and anxiety disorders in rheumatoid arthritis. Arthritis Care Res. 2019.
14. Khamseh ME, Baradaran HR, Javanbakht A, et al. Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran. BMC Psychiatry. 2011;11:61.

15. 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.
16. Lara MA, Navarrete L, Nieto L, et al. Prevalence and incidence of perinatal depression and depressive symptoms among Mexican women. J Affect Disord. 2015;175:18-24.
17. 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.
18. Marrie RA, Zhang L, Lix LM, et al. The validity and reliability of screening measures for depression and anxiety disorders in multiple sclerosis. Mult Scler Relat Dis. 2018;20:9-15.
19. Martin-Subero M, Kroenke K, Diez-Quevedo C, et al. Depression as measured by PHQ-9 versus clinical diagnosis as an independent predictor of long-term mortality in a prospective cohort of medical inpatients. Psychosom Med. 2017;79:273-282.
20. Osório FL, Vilela Mendes A, Crippa JA, et al. 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.
21. Osório FL, Carvalho AC, Fracalossi TA, et al. Are two items sufficient to screen for depression within the hospital context? Int J Psychiatry Med. 2012;44:141-148.
22. Patten SB, Burton JM, Fiest KM, et al. Validity of four screening scales for major depression in MS. Mult. Scler. 2015;21:1064-1071.
23. 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.
24. Prisnie JC, Fiest KM, Coutts SB, et al. Validating screening tools for depression in stroke and transient ischemic attack patients. Int J Psychiatry Med. 2016;51:262-277.
25. Shinn EH, Valentine A, Baum G, et al. Comparison of four brief depression screening instruments in ovarian cancer patients: Diagnostic accuracy using traditional versus alternative cutpoints. Gynecol Oncol. 2017;145:562-568.
26. Spangenberg L, Glaesmer H, Boecker M, et al. Differences in Patient Health Questionnaire and Aachen Depression Item Bank scores between tablet versus paper-and-pencil administration. Qual Life Res. 2015;24:3023-3032.
27. Wagner LI, Pugh SL, Small Jr W, et al. Screening for depression in cancer patients receiving radiotherapy: Feasibility and identification of effective tools in the NRG Oncology RTOG 0841 trial. Cancer. 2017;123:485-93.
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. Azah MN, Shah ME, Shaaban J, et al. Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. MedPulse. 2005;12:259-63.
30. De Man-van Ginkel JM, Hafsteinsdóttir T, Lindeman E, et al. An efficient way to detect poststroke depression by subsequent administration of a 9-item and a 2-item Patient Health Questionnaire. Stroke. 2012;43:854-56.
31. Fisher J, Rowe H, Wynter K, et al. Sex-informed, psychoeducational programme for couples to prevent postnatal common mental disorders among primiparous women: cluster randomised controlled trial. BMJ open. 2016;6:e009396.

32. Gelaye B, Tadesse MG, Williams MA, et al. Assessing validity of a depression screening instrument in the absence of a gold standard. Ann Epidemiol. 2014;24:527-31.
33. Grool AM, van der Graaf Y, Mali WP, et al. Location of cerebrovascular and degenerative changes, depressive symptoms and cognitive functioning in later life: the SMART-Medea study. J. Neurol. Neurosurg. Psychiatry. 2011;82:1093-1100.
34. Hahn D, Reuter K, Harter M. Screening for affective and anxiety disorders in medical patients - comparison of HADS, GHQ-12 and Brief-PHQ. Psychsoc Med. 2006;3.
35. Henkel V, Mergl R, Kohnen R, et al. Use of brief depression screening tools in primary care: consideration of heterogeneity in performance in different patient groups. Gen Hosp Psychiatry. 2004;26:190-98.
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. Kim DJ, Kim K, Lee HW, et al. Internet game addiction, depression, and escape from negative emotions in adulthood: a nationwide community sample of Korea. J Nerv Ment Dis. 2017;205:568-73.
38. Kohrt BA, Luitel NP, Acharya P, et al. Detection of depression in low resource settings: validation of the Patient Health Questionnaire (PHQ-9) and cultural concepts of distress in Nepal. BMC psychiatry. 2016;16:58.
39. Liu Y, Wang J. Validity of the patient health questionnaire-9 for DSM-IV major depressive disorder in a sample of Canadian working population. J Affect Disord. 2015;187:122-26.
40. 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;67:309-15.
41. 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;38:221-28.
42. Razykov I, Hudson M, Baron M, et al. Utility of the Patient Health Questionnaire-9 to assess suicide risk in patients with systemic sclerosis. Arth Care Res. 2013;65:753-58.
43. Zuithoff NP, Vergouwe Y, King M, et al. A clinical prediction rule for detecting major depressive disorder in primary care: the PREDICT-NL study. Fam Pract. 2009;26:241-50.
44. Akena D, Joska J, Obuku EA, et al. Sensitivity and specificity of clinician administered screening instruments in detecting depression among HIV-positive individuals in Uganda. AIDS Care. 2013;25:1245-52.
45. Baron EC, Davies T, Lund C. Validation of the 10-item centre for epidemiological studies depression scale (CES-D-10) in Zulu, Xhosa and Afrikaans populations in South Africa. BMC psychiatry. 2017;17:6.
46. Buji RI, Abdul Murad NA, Chan LF, et al. Suicidal ideation in systemic lupus erythematosus: NR2A gene polymorphism, clinical and psychosocial factors. Lupus. 2018;27:744-52.
47. 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;167:160-66.
48. Conway A, Sheridan J, Maddicks-Law J, et al. Accuracy of anxiety and depression screening tools in heart transplant recipients. Appl Nurs Res. 2016;32:177-81.

49. de la Torre AY, Oliva N, Echevarrieta PL, et al. Major depression in hospitalized Argentine general medical patients: Prevalence and risk factors. J Affect Disord. 2016;197:36-42.
50. Gholizadeh L, Shahmansouri N, Heydari M, et al. Assessment and detection of depression in patients with coronary artery disease: validation of the Persian version of the PHQ-9. Contemp Nurse. 2019;55:185-94.
51. Hantsoo L, Podcasy J, Sammel M, et al. Pregnancy and the acceptability of computer-based versus traditional mental health treatments. J Womens Health. 2017;26:1106-13.
52. Hides L, Lubman DI, Devlin H, et al. Reliability and validity of the Kessler 10 and Patient Health Questionnaire among injecting drug users. Aust N Z Psychiatry. 2007;41:166-68.
53. Hyphantis T, Kotsis K, Voulgari PV, et al. 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;63:1313-21.
54. 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;55:1950-59.
55. 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;35:592-97.
56. Janssen EP, Köhler S, Stehouwer CD, et al. The Patient Health Questionnaire-9 as a screening tool for depression in individuals with type 2 diabetes mellitus: The Maastricht study. J Am Geriatr Soc. 2016;64:e201-06.
57. Lamers F, Jonkers CC, Bosma H, et al. 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;61:679-87.
58. Levin-Aspenson HF, Watson D. Mode of administration effects in psychopathology assessment: Analyses of sex, age, and education differences in self-rated versus interview-based depression. Psychol. Assess. 2018;30:287.
59. Liu h, He YL, Miao JM, et al. Validity and reliability of the 12-item Short Form Health Survey in outpatients from traditional Chinese internal department. Chinese Mental Health Journal. 2016;30:6-12
60. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. BMC Psychiatry. 2008;8:46.
61. 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;101:952-60.
62. Muramatsu K, Miyaoka H, Kamijima K, et al. Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. Gen Hosp Psychiatry. 2018;52:64-9.
63. Nakku JE, Rathod SD, Kizza D, et al. Validity and diagnostic accuracy of the Luganda version of the 9-item and 2-item patient health questionnaire for detecting major depressive disorder in rural Uganda. GMH. 2016;3.
64. Paika V, Andreoulakis E, Ntountoulaki E, et al. The Greek-Orthodox version of the Brief Religious Coping (B-RCOPE) instrument: psychometric properties in three samples and

associations with mental disorders, suicidality, illness perceptions, and quality of life. Ann. Gen. Psychiatry. 2017;16:13.
65. 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.
66. Rancans E, Trapencieris M, Ivanovs R, et al. Validity of the PHQ-9 and PHQ-2 to screen for depression in nationwide primary care population in Latvia. Ann Gen Psychiatry. 2018;17:33.
67. 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;29:1533-43.
68. 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;29:417-24.
69. Sung SC, Low CC, Fung DS, et al. 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;5:249-58.
70. Suzuki K, Kumei S, Ohhira M, et al. Screening for major depressive disorder with the Patient Health Questionnaire (PHQ-9 and PHQ-2) in an outpatient clinic staffed by primary care physicians in Japan: a case control study. PloS one. 2015;10:e0119147.
71. van Heyningen T, Honikman S, Tomlinson M, et al. Comparison of mental health screening tools for detecting antenatal depression and anxiety disorders in South African women. PloS one. 2018;13:e0193697.
72. Volker D, Zijlstra-Vlasveld MC, Brouwers EP, et al. Validation of the patient health questionnaire-9 for major depressive disorder in the occupational health setting. J. Occup. Rehabil. 2016;26:237-44.
73. 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;151:660-66.
74. Alamri SH, Bari AI, Ali AT. Depression and associated factors in hospitalized elderly: a cross-sectional study in a Saudi teaching hospital. Ann. Saudi Med. 2017;37:122-129.
75. Bailer J, Kerstner T, Witthöft M, et al. Health anxiety and hypochondriasis in the light of DSM-5. Anxiety Stress Copin. 2016;29:219-39.
76. 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;32:271-83.
77. Brodey BB, Goodman SH, Baldasaro RE, et al. Development of the Perinatal Depression Inventory (PDI)-14 using item response theory: a comparison of the BDI-II, EPDS, PDI, and PHQ-9. Arch Womens Ment Health. 2016;19:307-16.
78. Chen S, Fang Y, Chiu H, et al. 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.
79. 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.

80. 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.
81. Irmak C, Altıntaş M. The resilience, attachment, coping, and psychopathology of battered women: comparison of sheltered versus in-home women. Ansdolu Psikiyatr De. 2017;18:561-70.
82. Lai BP, Tang AK, Lee DT, et al. Detecting postnatal depression in Chinese men: a comparison of three instruments. Psychiatry Res. 2010;180:80-85.
83. Limon FJ. Screening Latino Farmworkers for Depression in Primary Care. Dissertation Abstracts International: Section B: The Sciences and Engineering. 2016.
84. Liu ZW, Yu Y, Hu M, et al. PHQ-9 and PHQ-2 for screening depression in Chinese rural elderly. PloS one. 2016;11:e0151042.
85. Nacak Y, Morawa E, Tuffner D, et al. Insecure attachment style and cumulative traumatic life events in patients with somatoform pain disorder: A cross-sectional study. J Psychosom Res. 2017;103:77-82.
86. 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;138:343-51.
87. 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.
88. 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.
89. Vöhringer 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.
90. Watnick S, Wang PL, Demadura T, et al. Validation of 2 depression screening tools in dialysis patients. Am J Kid Dis. 2005;46:919-24.
91. 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.
92. 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.
93. Ikin JF, McKenzie DP, Gwini SM, et al. Major depression and depressive symptoms in Australian Gulf War veterans 20 years after the Gulf War. J Affect Disord. 2016;189:77-84.
94. Valencia-Garcia D, Bi X, Ayón C. Sensitivity and Specificity in Three Measures of Depression Among Mexican American Women. J Immigr Minor Healt. 2017;19:562-71.
95. Wang L, Lu K, Li J, et al. Value of patient health questionnaires (PHQ)-9 and PHQ-2 for screening depression disorders in cardiovascular outpatients. Zhonghua xin xue guan bing za zhi. 2015;43:428-31.
96. Choi SK, Boyle E, Burchell AN, et al. Validation of six short and ultra-short screening instruments for depression for people living with HIV in Ontario: results from the Ontario HIV treatment network cohort study. PLoS One. 2015;10:e0142706.

97. Griffith SD, Thompson NR, Rathore JS, et al. Incorporating patient-reported outcome measures into the electronic health record for research: application using the Patient Health Questionnaire (PHQ-9). Qual Life Res. 2015;24:295-303.
98. Persoons P, Luyckx K, Desloovere C, et al. 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;25:316-23.
99. 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;37:215-20.
100. Scott JD, Wang CC, Coppel E, et al. Diagnosis of depression in former injection drug users with chronic hepatitis C. J Clin Gastroenterol. 2011;45:462-67.
101. Seo JG, Park SP. Validation of the Patient Health Questionnaire-9 (PHQ-9) and PHQ-2 in patients with migraine. J Headache Pain. 2015;16:65.
102. 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.
103. 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;36:539-44.
104. Woldetensay YK, Belachew T, Tesfaye M, et al. Validation of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression in pregnant women: Afaan Oromo version. PloS one. 2018;13:e0191782.
105. Xiong N, Fritzsche K, Wei J, et al. Validation of patient health questionnaire (PHQ) for major depression in Chinese outpatients with multiple somatic symptoms: a multicenter cross-sectional study. J Affect Disord. 2015;174:636-43.
106. 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.
107. Bombardier CH, Kalpakjian CZ, Graves DE, et al. 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.
108. 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.
109. Fiest KM, Patten SB, Wiebe S, et al. Validating screening tools for depression in epilepsy. Epilepsia. 2014;55:1642-1650.
110. Gjerdingen D, Crow S, McGovern P, et al. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. Ann Fam Med. 2009;7:63-70.
111. 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. Natl Compr Canc Netw. 2015;13:1203-1211.
112. McGuire AW, Eastwood JA, Macabasco-O'Connell A, et al. Depression screening: utility of the Patient Health Questionnaire in patients with acute coronary syndrome. Am J Crit Care. 2013;22:12-19.

113. Richardson TM, He H, Podgorski C, et al. Screening depression aging services clients. Am J Geriatr Psychiatry. 2010;18:1116-1123.
114. 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 Oncol. 2013;15:122-129.
115. Sidebottom AC, Harrison PA, Godecker A, et al. Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. Arch Womens Ment Health. 2012;15:367-374.
116. Simning A, van Wijngaarden E, Fisher SG, et al. Mental healthcare need and service utilization in older adults living in public housing. Am J Geriatr Psychiatry. 2012;20:441-451.
117. 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.
118. Twist K, Stahl D, Amiel SA, et al. Comparison of depressive symptoms in type 2 diabetes using a two-stage survey design. Psychosom Med. 2013;75:791-797.
119. 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.
120. 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;8:348-53.
121. 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;134:266-71.
122. Kiely KM, Butterworth P. Validation of four measures of mental health against depression and generalized anxiety in a community based sample. Psychiatry Res. 2014;225:291-98.
123. 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;143:208-13.
124. 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;23:2014-17.