**Variability and bias in optimal cutoffs and accuracy estimates due to data-driven cutoff selection: Simulation study using data from 100 PHQ-9 accuracy studies**

Brooke Levis, PhDa,b, Parash Mani Bhandari, MSca, Dipika Neupane, MSca, Suiqiong Fan, MScPHa, Ying Sun, MPHa, Brett D. Thombs, PhDa-f,\*, Andrea Benedetti, PhDb,d,g,h,\*, and the DEPRESsion Screening Data (DEPRESSD) PHQ Groupi

aLady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; bDepartment of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada; cDepartment of Psychiatry, McGill University, Montréal, Québec, Canada; dDepartment of Medicine, McGill University, Montréal, Québec, Canada; eDepartment of Psychology, McGill University, Montréal, Québec, Canada; fBiomedical Ethics Unit, McGill University, Montréal, Québec, Canada; gRespiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada; hCentre for Outcomes Research& Evaluation, Research Institute of the McGill University Health Centre, Montréal, Québec, Canada; iMembers of the DEPRESSD PHQ Group: 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; Zelalem Negeri, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada; Mahrukh Imran, Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada; Danielle B. Rice, Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, 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; Pim Cuijpers, Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, the Netherlands; Simon Gilbody, Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK; John P. A. Ioannidis, Department of Medicine, Department of Epidemiology and Population Health, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA; Lorie A. Kloda, Library, Concordia University, Montréal, Québec, Canada; Scott B. Patten, Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada; Roy C. Ziegelstein, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; Daphna Harel, Department of Applied Statistics, Social Science, and Humanities, New York University, New York, New York, USA; Yemisi Takwoingi, Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK; Sarah Markham, Department of Biostatistics and Health Informatics, King's College London, London, UK; Sultan H. Alamri, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia; Dagmar Amtmann, Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA; Bruce Arroll, Department of General Practice and Primary Health Care, University of Auckland, New Zealand; 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, Durban, KwaZulu-Natal, South Africa; Charles H. Bombardier, Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, USA; Ryna Imma Buji, Department of Psychiatry, Hospital Mesra Bukit Padang, Sabah, Malaysia; Peter Butterworth, Centre for Research on Ageing, Health and Wellbeing, Research School of Population Health, The Australian National University, Canberra, Australia; Gregory Carter, Centre for Brain and Mental Health Research, University of Newcastle, New South Wales, Australia; 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; Kerrie Clover, Centre for Brain and Mental Health Research, University of Newcastle, New South Wales, Australia; Aaron Conway, Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, Toronto, Canada; Yeates Conwell, Department of Psychiatry, University of Rochester Medical Center, Rochester, New York, USA; Federico M. Daray, Institute of Pharmacology, School of Medicine, University of Buenos Aires, Argentina; Janneke M. de Man-van Ginkel, Leids Univesity Medical Center, Leiden, the Netherlands; Jesse R. Fann, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA; Felix H. Fischer, Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Germany; Sally Field, Perinatal Mental Health Project, Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa; Jane R. W. Fisher, Global and Women’s Health, Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia; Daniel Fung, Department of Developmental Psychiatry, Institute of Mental Health, Singapore; 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; Felicity Goodyear-Smith, Department of General Practice and Primary Health Care, University of Auckland, Auckland, New Zealand; Catherine G. Greeno, School of Social Work, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; Brian J. Hall, Center for Global Health Equity, 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, USA; Martin Härter, Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, 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, Cape Town, South Africa; Thomas Hyphantis, Department of Psychiatry, Faculty of Medicine, School of Health Sciences, University of Ioannina, Ioannina, Greece; Masatoshi Inagaki, Department of Psychiatry, Faculty of Medicine, Shimane University, Izumo, Shimane, Japan; Maria Iglesias-Gonzalez, Department of Psychiatry, Hospital Universitari Germans Trias i Pujol, Badalona, Spain; Hong Jin Jeon, Department of Psychiatry, Depression Center, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea; 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; Kim M. Kiely, School of Psychology, The University of New South Wales, and Neuroscience Research Australia (NeuRA), Sydney, Australia; Brandon A. Kohrt, Department of Psychiatry and Behavioral Sciences, The George Washington University, Washington, District of Columbia, USA; Yunxin Kwan, Department of Psychological Medicine, Tan Tock Seng Hospital, Singapore; Maria Asunción Lara, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz. San Lorenzo Huipulco, Tlalpan, Mexico; Holly F. Levin-Aspenson, Department of Psychology, University of North Texas, Texas, 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, Cape Town, South Africa; 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; Laura Marsh, Baylor College of Medicine, Houston and Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas, USA; Brian P. Marx, National Center for PTSD at VA Boston Healthcare System, Boston, Massachusetts, USA; Anthony McGuire, University of South Florida, College of Nursing, Tampa, Florida, 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, Brazil; Kumiko Muramatsu, Niigata Seiryo University Health Service Center, 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; Brian W. Pence, Department of Epidemiology, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Philippe Persoons, Department of Psycho-Pedagogic Psychiatry, Healthcare Group Sint-Kamillus, Broeders van Liefde, Bierbeek, Belgium; Inge Petersen, Centre for Rural Health, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, KwaZulu-Natal, South Africa; 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, Pennsylvania, USA; Terence J. Quinn, Institute of Cardiovascular & Medical Sciences, University of Glasgow, Glasgow, Scotland; Elmars Rancans, Department of Psychiatry and Narcology, Riga Stradins University, Riga, Latvia; Sujit D. Rathod, Department of Population Health, London School of Hygiene and Tropical Medicine, London, UK; Katrin Reuter, Group Practice for Psychotherapy and Psycho-oncology, Freiburg, Germany; Alasdair G. Rooney, Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburgh, Edinburgh, Scotland, UK; 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; Abbey Sidebottom, Allina Health, Minneapolis, Minnesota, USA; Adam Simning, Department of Psychiatry, University of Rochester Medical Center, Rochester, New York, USA; Lena Spangenberg, Department of Medical Psychology and Medical Sociology, University of Leipzig, Germany; Lesley Stafford, Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, 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, Melbourne, Victoria, Australia; Alyna Turner, School of Medicine and Public Health, University of Newcastle, New South Wales, Newcastle, Australia; Christina M. van der Feltz-Cornelis, Department of Health Sciences, HYMS, University of York, York, UK; Thandi van Heyningen, Institute for Security Studies, Pretoria, South Africa; Paul A. Vöhringer, Department of Psychiatry and Mental Health, Clinical Hospital, Universidad de Chile, Santiago, Chile; 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, Department of Community Health and Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada; David Watson, Department of Psychology, University of Notre Dame; Jennifer White, Department of Physiotherapy, School of Primary and Allied Health Care, Monash University, Melbourne, Australia; Mary A. Whooley, Departments of Medicine and of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA; Kirsty Winkley, Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, UK; Karen Wynter, School of Clinical Sciences, Monash University, Melbourne, Australia; Mitsuhiko Yamada, Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, 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.

\*Co-senior authors

**Corresponding author:**

Andrea Benedetti, PhD; Centre for Outcomes Research & Evaluation, Research Institute of the McGill University Health Centre, 5252 Boulevard de Maisonneuve, Montréal, Quebec, H4A 3S5, Canada; Tel (514) 934-1934 ext. 32161; E-mail: andrea.benedetti@mcgill.ca

**Word Count:** 2,937

**KEY POINTS**

**Question:** Does data-driven optimal cutoff selection in Patient Health Questionnaire-9 (PHQ-9) screening accuracy studies generate cutoffs that diverge from population cutoffs and overstate accuracy?

**Findings:** Optimal PHQ-9 cutoffs identified varied from ≥ 2 to ≥ 21 for samples with N = 100 and ≥ 5 to ≥ 11 for N = 1,000. Mean mis-estimation of sensitivity and specificity were 6.4 and -0.8 percentage points for samples of N = 100 and 1.8 and -0.6 percentage points for N = 1,000.

**Meaning:** In this simulation study, we illustrated how data-driven cutoff selection may mis-identify optimal cutoffs and exaggerate PHQ-9 screening accuracy.

**ABSTRACT**

**Importance:** Test accuracy studies often use small datasets to simultaneously select an “optimal” cutoff that maximizes test accuracy and generate accuracy estimates.

**Objective:** To conduct a simulation, or re-sampling, study with real participant data to evaluate the degree that data-driven optimal cutoff selection for screening for major depression with the Patient Health Questionnaire-9 (PHQ-9) may result in selection of optimal cutoffs that diverge from population optimal cutoffs and generate biased accuracy estimates.

**Design, Setting, and Participants:** We used data from an existing individual participant data meta-analysis database on PHQ-9 screening accuracy to represent a hypothetical population. Studies in the database compared participant PHQ-9 scores with major depression classification. From the population, we randomly sampled 1,000 studies each of 100, 200, 500, and 1,000 participants.

**Main outcomes and measures:** For the full population and each simulated study, an optimal cutoff was selected by maximizing Youden’s J. Accuracy estimates for optimal cutoffs in simulated studies were compared to accuracy in the full population.

**Results:** The population included 44,503 participants (4,541 with major depression) from 100 primary studies. The population optimal cutoff was ≥ 8. Optimal cutoffs in simulated studies ranged from 2-21 in samples of 100, 3-14 in samples of 200, 4-13 in samples of 500, and 5-11 in samples of 1,000 participants. The percentage of simulated studies that identified the true optimal cutoff of ≥ 8 was 17% for samples of 100, 22% for samples of 200, 26% for samples of 500, and 33% for samples of 1,000. Compared to estimates for cutoff ≥ 8 in the population, sensitivity was overestimated by 6.4 percentage points, 4.9 percentage points, 2.2 percentage points, and 1.8 percentage points in samples of 100, 200, 500, and 1,000 participants. Specificity was within 1 percentage point across sample sizes.

**Conclusions and relevance:** In this simulation study, optimal cutoffs and accuracy estimates differed substantially from population values when data-driven methods were used to simultaneously identify an optimal cutoff and estimate accuracy. Clinicians should use cutoffs generated from well-conducted meta-analyses and adopt population-specific cutoffs that diverge from general cutoffs only if from large studies or replicated in multiple studies.

**Registration:** PROSPERO (CRD42014010673)

**Keywords:** data-driven analysis, depression, individual participant data meta-analysis, methods, screening, simulation study

**Running Title:** Data-driven cutoff selection with the PHQ-9

**Funding:** Canadian Institutes of Health Research (KRS-134297, PCG-155468, PJT-162206).

1. **INTRODUCTION**

Studies on depression screening tool accuracy often use data-driven approaches and small samples and numbers of depression cases to simultaneously determine an ‘optimal’ cutoff and estimate accuracy [1-3]. A recent review of 172 studies found that median sample size was 194 and median number of depression cases approximately 20 [1]. Seventy-six percent of included studies identified an ‘optimal’ cutoff that diverged from a standard cutoff, and authors of 40% of those studies recommended that their optimal cutoff, rather than the standard cutoff, be used in their population [1].

Previous studies on data-driven selection of test cutoffs have reported that these methods produce overly optimistic accuracy estimates, especially in small samples [4-8]. However, most of these studies used simulated datasets based on hypothetical test score distributions rather than real participant data. We previously analyzed Edinburgh Postnatal Depression Scale (EPDS) data from 13,255 participants and found that in 1,000 simulated or “re-sampled” studies, the cutoff maximizing Youden’s J (sensitivity + specificity – 1) [9] ranged from ≥ 5 to ≥ 17 with re-sampled studies of N = 100 and ≥ 8 to ≥ 13 with N = 1,000. Mean sensitivity overestimation was 7 percentage points (pp) for N = 100 versus 1pp for N = 1,000, while specificity was underestimated by 1pp across sample sizes [8].

The standard cutoff traditionally used to screen for major depression with the Patient Health Questionnaire-9 (PHQ-9) is ≥ 10 [10-14]. An individual participant data meta-analysis (IPDMA) of 100 primary studies (44,503 participants, 4,541 major depression cases) confirmed that a cutoff of ≥ 10 maximized combined sensitivity and specificity in studies that used a “gold standard” semi-structured interview reference standard, although the optimal cutoff was ≥ 8 when fully structured interviews, designed for lay administration, were used [15, 16].

Many primary studies of PHQ-9 accuracy emphasize results from data-driven optimal cutoffs. [1].The degree that accuracy is overestimated when data-driven cutoffs are used for the PHQ-9, however, is not known. The objective of the present study was to determine the degree that using data-driven methods to simultaneously select an optimal PHQ-9 cutoff and estimate accuracy may bias estimates. We estimated, across different sample sizes, the degree that data-driven cutoff selection results in (1) sample-specific optimal cutoffs that differ from the population-level optimal cutoff and (2) biased accuracy estimates. In addition, for comparison, we (3) estimated accuracy using the population-level optimal cutoff in individual re-sampled studies and compared to population accuracy.

1. **MATERIALS & METHODS**

We used data from an IPDMA of PHQ-9 diagnostic accuracy to represent a hypothetical “population” from which studies of different sizes could be re-sampled [16]. The main IPDMA was registered in PROSPERO (CRD42014010673), and a protocol was published [19]. A protocol for the present study was published in the Open Science Framework repository prior to initiation [20]. Details on the methodology used to identify, obtain, and synthesize the data included in the present study are provided in eMethods1 and eMethods2. We used a similar methodological approach as our EPDS simulation study [8]. Because of the overlap of methods in the present study and previous studies, we described the methods similarly and followed reporting guidance from the Text Recycling Research Project [21].

**2.1 Data analysis**

For the purposes of the present study, we used our IPDMA dataset to represent a hypothetical population and defined population sensitivity and specificity values for PHQ-9 cutoffs to be those estimated in the hypothetical population. In our main IPDMA on PHQ-9 accuracy, we accounted for clustering of observations within each study, and we applied sampling weights to account for imbalances in participant samples when, for instance, all participants with positive screens but only a random portion of those with negative screens were administered a diagnostic interview. In the present study, we ignored clustering and sampling weights to have a defined population from which we could draw samples that represented simulated primary studies and to be able to analyze the population data and simulated primary study data with the same analytical approach. In addition, in the main IPDMA, we stratified included studies by reference standard type because previous studies have shown that different types of diagnostic interviews classify major depression differently [22-23]. However, in main analyses of the present study, we did not stratify studies by reference standard because we were not evaluating the true screening accuracy of the PHQ-9, and combining included studies that used different reference standards allowed us to have a single hypothetical population for simulation. As a result, this procedure produced accuracy estimates that differed from those reported in the published IPDMA [16]. In the present study, we calculated the population optimal cutoff that maximized Youden’s J in the full IPDMA dataset, which was ≥ 8.

First, we described the individual primary studies included in the IPDMA dataset in terms of sample size, number of major depression cases, and optimal cutoff (based on maximizing Youden’s J). If there was a tie in maximum Youden’s J between multiple cutoffs, we randomly selected one of the cutoffs. We chose to use Youden’s J because it is by far the most commonly used method for selecting optimal cutoffs in depression screening accuracy studies, and our re-sampling study aimed to reflect current research practices [1].

Second, from the IPDMA dataset, we sampled with replacement to generate 1,000 randomly sampled studies each of 100, 200, 500, and 1,000 participants to mimic what would occur in primary studies that use samples of these sizes. For each study, we defined the sample-specific optimal cutoff as the cutoff that maximized Youden’s J with random selection in case of ties. For each sample size, across the 1,000 samples, we (1) graphically illustrated the variability in sample-specific optimal cutoffs and their accuracy estimates; and (2) estimated the mean difference in sensitivity and specificity estimates at the sample-specific optimal cutoffs and at a cutoff of ≥ 8, compared to sensitivity and specificity estimates for a cutoff of ≥ 8 in the population. In additional analyses, we stratified results by optimal cutoff value.

Random selection of participants in simulated samples and averaging sensitivity and specificity across 1,000 samples for each sample size was done to balance other possible sources of divergent accuracy, such as reference standards or individual participant characteristics. Nonetheless, in sensitivity analyses, we repeated the simulation process including only studies that used the semi-structured Structured Clinical Interview for DSM (SCID) as the reference standard.

For all analyses, sensitivity and specificity were estimated using 2 x 2 table counts. Analyses were done using R version 4.2.2.

**2.2 Ethical Approval**

The Research Ethics Committee of the Jewish General Hospital (Montreal, Quebec) determined that research ethics approval was not required, since the study involved IPDMA of anonymized previously collected data. For each included dataset, we confirmed that the original study received ethics approval and that participants provided informed consent.

**3. RESULTS**

**3.1 Description of the 100 included primary studies**

The full IPDMA database included 100 primary studies with 44,503 participants (4,541 major depression cases, 10%), which constituted the “population” for the present study. See eTable1 for primary study characteristics. In the 100 included primary studies, mean sample size was 445 participants (median = 194) and mean number of major depression cases was 45 (median = 29). Study-specific optimal cutoffs ranged from ≥ 3 to ≥ 18 (median = ≥ 10). Frequencies of PHQ-9 scores for cases and non-cases are shown in eTable2 with histograms in eFigure1. PHQ-9 scores were normally distributed (mean and median = 13) among cases and right-skewed among non-cases (mean = 4, median = 3). In the full database population, unweighted sensitivity and specificity for PHQ-9 ≥ 8 were 80.4% and 82.0%.

**3.2 Variability of sample-specific optimal cutoffs in simulated samples**

Figure 1 shows the variability of sample-specific optimal cutoffs from 1,000 re-sampled studies of 100, 200, 500 and 1,000 participants. As sample size increased, the variability in sample-specific optimal cutoffs decreased. In studies of 100 participants, study-specific optimal cutoffs ranged from ≥ 2 to ≥ 21; 17% of studies had an optimal cutoff of ≥ 8 and 45% had an optimal cutoff between ≥ 7 and ≥ 9. When sample size increased to 1,000 participants, the range of optimal cutoffs was ≥ 5 to ≥ 11; 33% of samples had an optimal cutoff of ≥ 8, and 79% had an optimal cutoff between ≥ 7 and ≥ 9.

**3.3 Bias analyses in simulated samples**

As shown in Figure 2, overestimation of sensitivity estimates for sample-specific optimal cutoffs decreased with increasing sample size, whereas specificity estimates remained within 1pp across sample sizes. Precision of both sensitivity and specificity estimates increased with sample size. As shown in Table 1, compared to accuracy estimates for a cutoff of ≥ 8 in the full database, study-specific optimal cutoffs in samples of 100 participants overestimated sensitivity by a mean of 6.4pp (95% confidence interval [CI] 5.7 to 7.1) and overestimated specificity by 0.6pp (95% CI 0.0 to 1.2). When sample size increased to 1,000, study-specific optimal cutoffs overestimated sensitivity by 1.8pp (95% CI 1.5 to 2.1) and underestimated specificity by 0.6pp (95% CI -1.0 to -0.3).

As shown in Table 1 and Figure 3, when each re-sampled study used a pre-specified cutoff of ≥ 8, mean sample-specific sensitivity and specificity values were similar to those in the population for all sample sizes.

As shown in eTable3, across sample sizes, bias in estimates increased as the sample-specific optimal cutoff diverged from ≥ 8. When the sample-specific optimal cutoff was lower than ≥ 8, specificity was underestimated, whereas when the sample-specific optimal cutoff was higher than ≥ 8, specificity was overestimated. The opposite pattern was seen for sensitivity, although there was a shift in values, given that even when the sample-specific cutoff was exactly ≥ 8, sensitivity was, on average, overestimated.

**3.4 Sensitivity Analyses**

As shown in eTable4 and eTable5, variability in sample-specific optimal cutoffs and bias in sensitivity and specificity were similar to the main results when only studies that used the SCID reference standard were included.

1. **DISCUSSION**

**4.1 Principal Findings**

This was the first study to assess bias in PHQ-9 accuracy estimates due to data-driven optimal cutoff selection. Strengths included the use of a large sample and real participant data.

The main finding of this simulation, or re-sampling, study was that data-driven optimal PHQ-9 cutoffs often differed from the population optimal cutoff, sometimes substantially, and produced biased accuracy estimates. As sample size increased from 100 to 1,000 participants, variability in optimal cutoffs decreased from a range of ≥ 2 to ≥ 21 to a range of ≥ 5 to ≥ 11 and overestimation in sensitivity compared to the population value decreased from 6.4pp to 1.8pp, while specificity remained within 1pp. The magnitude and direction of bias differed depending on how far the sample-specific optimal cutoff was from the population optimal cutoff of ≥ 8. When a pre-defined cutoff of ≥ 8 was used in re-sampled studies, mean accuracy estimates were consistent with overall population estimates.

**4.2 Comparison with Other Studies**

Previous distribution-based simulation studies have found that data-driven cutoff selection in small samples results in exaggerated accuracy estimates [4-7]. Most studies on depression screening tool accuracy have small sample sizes and numbers of depression cases. Individual studies often report results from one, several, or many cutoffs, with the result being a wide range of optimal cutoffs and accuracy estimates across studies [1-3]. Many authors conclude that sample characteristics influence accuracy and that different optimal cutoffs are needed for particular population subgroups. Results from the present study and the previous simulation study on the EPDS [8] suggest that variability in optimal cutoffs and accuracy estimates often occurs due to chance and imprecision in small samples even when all samples are drawn from the same population. The findings that data-driven methods and small samples may explain divergent results across studies are consistent with the results of several large IPDMAs [15,16,24,25],which found that there were no substantive differences in depression screening tool accuracy based on participant characteristics. Additionally, the finding in the present study that accuracy estimates were similar between the full population and re-sampled studies when the same cutoff was used underlines that divergences can be attributed to data-driven methods and sample size rather than characteristics of participants in each sample.

Our finding that there were larger biases in sensitivity compared to specificity is not surprising, given that most studies have many fewer participants with depression than without. In addition, PHQ-9 scores among cases were normally distributed, whereas scores among non-cases were heavily right-skewed. Similar results were seen in the simulation study of EPDS data, which found that overestimation of sensitivity reduced from 7pp in samples of 100 participants to 1pp in samples of 1,000 participants, while specificity was underestimated by 1pp across sample sizes [8].These findings suggest that data-driven methods for cutoff selection can allow for substantial sensitivity gains with only minor costs to specificity, although at the individual study level, sensitivity can be either over- or underestimated.

**4.3 Implications**

Clinicians and policy makers who make decisions regarding depression screening should be aware that optimal cutoffs for the PHQ-9 and other depression screening tools identified in small single studies should be interpreted cautiously. Decisions regarding what cutoffs to use should ideally be based on large, well-conducted meta-analyses, or based on multiple validations in studies with adequate sample size for desired precision levels. In addition, clinicians may want to prioritize either sensitivity or specificity in different clinical settings, rather than consider them equally as is the case when selecting cutoffs based on Youden’s J, and select higher or lower cutoffs, depending on health and resource priorities [26]. The optimal cutoff of ≥ 8, which was identified in our hypothetical population was not derived using methods that account for clustering and sample weights and may reflect that participants from studies that used different reference standards were combined. Cutoffs and accuracy estimates from the main PHQ-9 IPDMA should be used clinically [16]. Depression screening questionnaires are not intended to establish clinical diagnoses but can be used for screening followed by clinical evaluation of those who screen positive. Whether screening should occur in practice requires evidence from clinical trials of benefit, which has not been established [27].

Although the Standards for Reporting of Diagnostic Accuracy Studies (STARD) reporting guideline recommends a priori sample size calculations [28], most depression screening tool accuracy studies do not conduct such calculations [2,3]. Researchers conducting primary studies on accuracy should conduct sample size calculations prior to recruitment to ensure the inclusion of sufficient numbers of both cases and non-cases for desired precision levels in accuracy estimates [29]. In addition, selective cutoff reporting bias occurs when researchers select which cutoffs to report accuracy results for in their individual studies based on the relative accuracy at those cutoffs in their sample (e.g., reporting accuracy estimates for cutoffs that maximize Youden’s J, but not for other cutoffs) [17,18]. Selective cutoff reporting bias has been found to underestimate sensitivity for cutoffs below a clearly defined standard cutoff and overestimate sensitivity for cutoffs above the standard [17,18]. Since summary accuracy estimates for a pre-defined cutoff do not tend to be biased, researchers should report accuracy estimates for all possible cutoffs rather than just those that are optimal in a given study or cutoffs close to the optimal cutoff [17,18]. Additionally, statistical methods for estimating cutoffs and “out-of-sample” performance, such as smoothing based on kernel estimation and bootstrapping, should be considered [30].

Beyond variability in accuracy estimates, researchers should also consider variability in the optimal cutoff that may be identified in individual studies. It is possible that authors of primary studies could use statistical methods to estimate uncertainty around optimal cutoffs in their individual studies, for instance via confidence intervals [31,32], and use internal validation methods (e.g., bootstrapping) to adjust for bias due to optimism [30,33]. Further work to test and demonstrate such methods for the purpose of mental health screening is needed.

**4.4 Limitations**

A limitation to consider is that we did not include datasets from recently published studies on PHQ-9 accuracy; however, we do not expect that the inclusion of more recent studies would alter results, given that newer studies would likely be of similar sample sizes and heterogeneity. We included data from 100 primary studies, and we believe that the dataset used for this study adequately represents a hypothetical population for simulation purposes. A second possible limitation is that we used only Youden's J to select optimal cutoffs. It is by far the most common method in depression screening accuracy studies [1] and performs similarly to other indices (e.g., the Euclidean distance) [34], but it is known to be unreliable and prone to overestimation. It is possible that results could differ slightly for an alternative method.

1. **CONCLUSIONS**

In summary, using samples with small numbers of participants and cases to simultaneously identify an optimal cutoff and estimate its accuracy results in optimal cutoffs that vary widely from study to study and exaggerated accuracy estimates. Variability in optimal cutoffs and the extent of sensitivity exaggeration reduces as sample size increases. Researchers should conduct a priori sample size calculations to ensure the inclusion of sufficient numbers of both cases and non-cases in diagnostic accuracy studies, report accuracy estimates for all cutoffs rather than just study-specific optimal cutoffs, and avoid making recommendations about optimal cutoffs and accuracy based on small single studies. They should consider using statistical methods that improve optimal cutoff identification and estimation of accuracy outside of the study sample. Users of diagnostic accuracy evidence, including researchers, clinicians, and policy makers, should evaluate studies of PHQ-9 accuracy with caution and ensure that recommendations regarding cutoffs are based on adequately powered and analyzed research studies or well-conducted meta-analyses.

**Contributions:** BLevis, BDT, and ABenedetti, and were responsible for the study conception and design. BLevis, PMB, DN, SFan, YS, and BDT contributed to data extraction, coding, evaluation of included studies, and data synthesis. BLevis, BDT, and ABenedetti contributed to data analysis and interpretation. BLevis drafted the manuscript, with the support of BDT and ABenedetti.

Members of the DEPRESSD PHQ Group contributed:

To data extraction, coding, and synthesis: CH, YW, AK, ZN, MImran, DBR, KER, MA, AWL. Via the design and conduct of database searches: JB, LAK. As members of the DEPRESSD Steering Committee, including conception and oversight of collaboration: PC, SG, JPAI, SBP, RCZ, YT. As a knowledge user consultant: SM. By contributing included datasets: SBP, SHA, DA, BA, LA, HRB, ABeraldi, CNB, ABhana, CHB, RIB, PB, GC, MHC, JCNC, LFC, DC, KC, AC, YC, FMD, JMdMvG, JRF, FHF, SField, JRWF, DF, BG, LJG, FGS, EPG, CGG, BJH, LHantsoo, MH, LHides, SEH, SH, TH, MInagaki, MIG, HJJ, NJ, MEK, KMK, BAK, YK, MLara, HFLA, SIL, MLotrakul, SRL, BLöwe, NPL, CL, RAM, LM, PBM, AM, SMS, TNM, KM, JEMN, LN, FLO, BWP, PP, IP, AP, SLP, TJQ, ER, SDR, KR, AGR, ISS, MTS, JS, EHS, ASidebottom, ASimning, LSpangenberg, LStafford, SCS, KS, PLLT, MTR, TDT, AT, CMvdFC, TvH, PAV, LIW, JLW, DW, JW MAW, KWinkley, KWynter, MY, QZZ, YZ.

All authors, including group authors, provided a critical review and approved the final manuscript. BDT and ABenedetti contributed equally as co-senior authors and are the guarantors; they had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

**Patient and Public Involvement:** No patients were involved in developing the research question, outcome measures, or study design. Since study inception, Dr. Sarah Markham joined the DEPRESSD Group as a patient collaborator. She reviewed the draft manuscript for this study. There are no plans to disseminate the results of the research to patients except through journal publication.

**Funding:** This study was funded by the Canadian Institutes of Health Research (CIHR; KRS-134297, PCG-155468). Dr. Levis was supported by a Fonds de recherche du Québec - Santé (FRQS) Postdoctoral Training Fellowship. Mr. Bhandari was supported by a studentship from the Research Institute of the McGill University Health Centre. Ms. Neupane was supported by a G.R. Caverhill Fellowship from the Faculty of Medicine, McGill University. Dr. Thombs was supported by a Tier 1 Canada Research Chair. Dr. Benedetti was supported by a Fonds de recherche du Québec - Santé (FRQS) researcher salary award.

Dr. Wu was supported by a FRQS Postdoctoral Training Fellowship. The primary studies by Fiest et al., Patten et al., Amoozegar 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. Jetté was supported by a Canada Research Chair in Neurological Health Services Research and an AIHS Population Health Investigator Award. She is Icahn School of Medicine at Mount Sinai Bludhorn Professor of International Medicine. 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). Collection of data for the study by Arroll et al. was supported by a project grant from the Health Research Council of New Zealand. Data collection for the study by Ayalon et al. was supported from a grant from Lundbeck International. The primary study by Khamseh et al. was supported by a grant (M-288) from Tehran University of Medical Sciences. The primary 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 Bombardier et al. was supported by the Department of Education, National Institute on Disability and Rehabilitation Research, Spinal Cord Injury Model Systems: University of Washington (grant no. H133N060033), Baylor College of Medicine (grant no. H133N060003), and University of Michigan (grant no. H133N060032). 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 Kiely et al. was supported by National Health and Medical Research Council (grant number 1002160) and Safe Work Australia. Dr. Butterworth was supported by Australian Research Council Future Fellowship FT130101444. Collection of data for the primary study by Zhang et al. was supported by the European Foundation for Study of Diabetes, the Chinese Diabetes Society, Lilly Foundation, Asia Diabetes Foundation and Liao Wun Yuk Diabetes Memorial Fund. The primary study by Chibanda et al. was supported by a grant from Grand Challenges Canada (0087-04). Dr. Cholera was supported by a United States National Institute of Mental Health (NIMH) grant (5F30MH096664), and the United States National Institutes of Health (NIH) Office of the Director, Fogarty International Center, Office of AIDS Research, National Cancer Center, National Heart, Blood, and Lung Institute, and the NIH Office of Research for Women’s Health through the Fogarty Global Health Fellows Program Consortium (1R25TW00934001) and the American Recovery and Reinvestment Act. 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). Dr. Conwell received support from NIMH (R24MH071604) and the Centers for Disease Control and Prevention (R49 CE002093). 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). Collection of data for the primary study by Delgadillo et al. was supported by grant from St. Anne’s Community Services, Leeds, UK. 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). Collection of data for the primary study by Fann et al. was supported by grant RO1 HD39415 from the US National Center for Medical Rehabilitation Research. The primary study by Fischer et al. was funded by the German Federal Ministry of Education and Research (01GY1150). 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). The primary study by Eack et al. was funded by the NIMH (R24 MH56858). Collection of data for the primary study by Hobfoll et al. was made possible in part from grants from NIMH (RO1 MH073687) and the Ohio Board of Regents. Dr. Hall received support from a grant awarded by the Research and Development Administration Office, University of Macau (MYRG2015-00109-FSS). 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. 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, FRQS, the Canadian Arthritis Network, and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montréal, QC. 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 Twist et al. was funded by the UK National Institute for Health Research under its Programme Grants for Applied Research Programme (grant reference number RP-PG-0606-1142). The 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. Kiely was supported by funding from an Australian National Health and Medical Research Council fellowship (grant number 1088313). 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).502: 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. 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. Collection of data for the primary study by Williams et al. was supported by a NIMH grant to Dr. Marsh (RO1-MH069666). 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). Collection of primary data for the study by Pence et al. was provided by NIMH (R34MH084673). 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.). The primary study by Rooney et al. was funded by the UK National Health Service Lothian Neuro-Oncology Endowment Fund. 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 Sidebottom et al. was funded by a grant from the United States Department of Health and Human Services, Health Resources and Services Administration (grant number R40MC07840). Simning et al.’s research was supported in part by grants from the NIH (T32 GM07356), Agency for Healthcare Research and Quality (R36 HS018246), NIMH (R24 MH071604), and the National Center for Research Resources (TL1 RR024135). 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. Collection of data for the study by Turner et al. (2012) was funded by a bequest from Jennie Thomas through the Hunter Medical Research Institute. The study by van Steenbergen-Weijenburg et al. was funded by Innovatiefonds Zorgverzekeraars. 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. Dr Vöhringer was supported by the Fund for Innovation and Competitiveness of the Chilean Ministry of Economy, Development and Tourism, through the Millennium Scientific Initiative (grant number IS130005). The primary study by Liu et al. (2015) was supported by CIHR (MOP-114970). The primary study by Wang et al. was supported by grants from the Shanghai Health Bureau (Grant number XBR2011015) and from the Ministry of Science and Technology project (2009BAI77B08, 2012BAI01B07). The primary study by Thombs et al. was done with data from the Heart and Soul Study. The Heart and Soul Study was funded by the Department of Veterans Epidemiology Merit Review Program, the Department of Veterans Affairs Health Services Research and Development service, the National Heart Lung and Blood Institute (R01 HL079235), the American Federation for Aging Research, the Robert Wood Johnson Foundation, and the Ischemia Research and Education Foundation. The primary study by 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).

**Declaration of Competing Interests:** All authors have completed the ICJME uniform disclosure form at www.icmje.org/coi\_disclosure.pdfand declare: no support from any organization for the submitted work; no financial relationships with any organizations 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. 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.

**Ethical Approval:** The Research Ethics Committee of the Jewish General Hospital declared that research ethics approval was not required, since the study involved IPDMA of anonymized previously collected data. However, for each included dataset, we confirmed that the original study received ethics approval and that participants provided informed consent.

**Data Statement:** Requests to access data should be made to the corresponding author.

**REFERENCES**

1. Brehaut E, Neupane D, Levis B, et al. 'Optimal' cutoff selection in studies of depression screening tool accuracy using the PHQ-9, EPDS, or HADS-D: a meta-research study. *Int J Methods Psychiatr Res* 2022:e1956. doi:10.1002/mpr.1956
2. Thombs BD, Rice DB. Sample sizes and precision of estimates of sensitivity and specificity from primary studies on the diagnostic accuracy of depression screening tools: a survey of recently published studies*. Int J Methods Psychiatr Res* 2016;25:145-52. doi:10.1002/mpr.1504
3. Nassar EL, Levis B, Neyer MA, et al. Sample size and precision of estimates in studies of depression screening tool accuracy: a meta-research review of studies published in 2018-2021. *Int J Methods Psychiatr Res* 2022;31(2):e1910. doi:10.1002/mpr.1910
4. Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin Chem* 1986;32:1341-6. [doi:10.1093/clinchem/32.7.1341](https://doi.org/10.1093/clinchem/32.7.1341)
5. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research*. J Clin Epidemiol* 2006;59(8):798-801. doi:10.1016/j.jclinepi.2005.11.025
6. Leeflang MM, Moons KG, Reitsma JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions*. Clin Chem* 2008;54(4):729-37. doi:10.1373/clinchem.2007.096032
7. Hirschfeld G, do Brasil PE. A simulation study into the performance of "optimal" diagnostic thresholds in the population: "large" effect sizes are not enough*. J Clin Epidemiol* 2014;67:449-53. [doi:10.1016/j.jclinepi.2013.07.018](https://doi.org/10.1016/j.jclinepi.2013.07.018)
8. Bhandari PM, Levis B, Neupane D, et al. Data-driven methods distort optimal cutoffs and accuracy estimates of depression screening tools: a simulation study using individual participant data. *J Clin Epidemiol* 2021;137:137-147. doi:10.1016/j.jclinepi.2021.03.031
9. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32-5. doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3
10. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606-13. [doi:10.1046/j.1525-1497.2001.016009606.x](https://psycnet.apa.org/doi/10.1046/j.1525-1497.2001.016009606.x)
11. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 2002;32:1-7. [doi:10.3928/0048-5713-20020901-06](https://psycnet.apa.org/doi/10.3928/0048-5713-20020901-06)
12. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA* 1999;282:1737-44. doi:10.1001/jama.282.18.1737
13. Wittkampf KA, Naeije L, Schene AH, et al. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007;29:388-95. doi:10.1016/j.genhosppsych.2007.06.004
14. Gilbody S, Richards D, Brealey S, et al. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007;22:1596-602. doi: 10.1007/s11606-007-0333-y
15. Levis B, Benedetti A, Thombs BD, et al. The diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: an individual participant data meta-analysis. *BMJ.* 2019;365:I1476. [doi:10.1136/bmj.l1476](https://doi.org/10.1136/bmj.l1476)
16. Negeri ZF, Levis B, Sun Y, et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. *BMJ* 2021;375:n2183. [doi:10.1136/bmj.n2183](https://doi.org/10.1136/bmj.n2183)
17. Levis B, Benedetti A, Levis AW, et al. Selective cutoff reporting in studies of diagnostic test accuracy: a comparison of traditional and individual patient data meta-analyses of the Patient Health Questionnaire-9 depression screening tool. *Am J Epidemiol* 2017;185:954-64. doi:10.1093/aje/kww191
18. Neupane D, Levis B, Bhandari PM, et al. Selective cutoff reporting in studies of the accuracy of the Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale: comparison of results based on published cutoffs versus all cutoffs using individual participant data meta-analysis. *Int J Methods Psychiatr Res* 2021;30:e1873. doi:10.1002/mpr.1873
19. Thombs BD, Benedetti A, Kloda LA, et al. The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Syst Rev* 2014;3:124. doi:10.1186/2046-4053-3-124
20. Levis B, Bhandari PM, Benedetti A, Thombs BD, DEPRESSD Collaboration. Evaluation of bias in diagnostic accuracy estimates due to data-driven cutoff selection: protocol for a simulation study using individual participant data from 58 studies on the diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9). March 2019. <https://osf.io/pg2yc/>. Accessed April 28, 2024.
21. Hall S, Moskovitz C, Pemberton M. Text Recycling Research Project. Best practices for researchers. V1.1 April, [cited 2021; <https://textrecycling.org/resources/best-practices-for-researchers/>. Accessed September 15, 2023.
22. Levis B, Benedetti A, Riehm KE, et al. Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. *Br J Psychiatry J Ment Sci* 2018;212:377-85. doi:10.1192/bjp.2018.54
23. Wu Y, Levis B, Ioannidis JPA, et al. 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* 2021;90:28-40. [doi:10.1159/000509283](https://doi.org/10.1159/000509283)
24. Levis B, Negeri Z, Sun Y, et al. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. *BMJ* 2020;371:m4022. [doi:10.1136/bmj.m4022](https://doi.org/10.1136/bmj.m4022)
25. Wu Y, Levis B, Sun Y, et al. Accuracy of the Hospital Anxiety and Depression Scale Depression subscale (HADS-D) to screen for major depression: systematic review and individual participant data meta-analysis. *BMJ* 2021;373:n972. doi:10.1136/bmj.n972
26. Using the PHQ-9 to screen for depression: A practice-based perspective. 2019. <http://depressionscreening100.com/phq/>. Accessed April 28, 2024.
27. Thombs BD, Markham S, Rice DB, Ziegelstein RC. Does depression screening in primary care improve mental health outcomes? *BMJ* 2021;374:n1661.
28. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies*. BMJ* 2015;351:h5527. [doi:10.1136/bmj.h5527](https://doi.org/10.1136/bmj.h5527)
29. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;58:859-62. doi:10.1016/j.jclinepi.2004.12.009
30. Thiele C, Hirshfeld G. cutpointr: Improved estimation and validation of optimal cutpoints in R. *J Stat Softw* 2021;98:1-27. doi: 10.18637/jss.v098.i11
31. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J* 2005;47:458-72. doi:10.1002/bimj.200410135
32. Schisterman EF, Perkins N. Confidence intervals for the Youden Index and corresponding optimal cut-point. *Commun Stat Simul Comput* 2007;36:549-63. doi:10.1080/0361091070212181
33. Smith GC, Seaman SR, Wood AM, et al. Correcting for optimistic prediction in small data sets. *Am J Epidemiol* 2014;180:318-24. doi:10.1093/aje/kwu140
34. Hajian-Tilaki K. The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. *Stat Methods Med Res* 2018;27:2374-83. doi:10.1177/0962280216680383

**Figure 1. Variability of data-driven optimal cutoffs in 1,000 samples of size 100, 200, 500, and 1,000**

**Figure 2. Variability in accuracy estimates of the optimal cutoffs in 1,000 samples of size 100, 200, 500, and 1,000, compared to accuracy estimates from a cutoff of ≥ 8 in the population**

Edges of boxes represent the 25th and 75th percentiles; mid-lines represent the medians

Dotted horizontal line represents the accuracy of the *true population optimal cutoff* in the full PHQ-9 IPDMA dataset (cutoff of ≥ 8; sensitivity = 80.4%, specificity = 82.0%)

**Figure 3. Variability in accuracy estimates of a cutoff of ≥ 8 in 1,000 samples of size 100, 200, 500, and 1,000, compared to accuracy estimates from a cutoff of ≥ 8 in the population**

Edges of boxes represent the 25th and 75th percentiles; mid-lines represent the medians

Dotted horizontal line represents the accuracy of the *true population optimal cutoff* in the full PHQ-9 IPDMA dataset (cutoff of ≥ 8; sensitivity = 80.4%, specificity = 82.0%)

**Table 1. Mean bias of accuracy estimates (with 95% confidence intervals) for 1,000 samples of size 100, 200, 500, and 1,000**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sample size = 100 | Sample size = 200 | Sample size = 500 | Sample size = 1,000 |
| **Sensitivity** | **Specificity** | **Sensitivity** | **Specificity** | **Sensitivity** | **Specificity** | **Sensitivity** | **Specificity** |
| Sample-specific optimal cutoffa – Population optimal cutoff of ≥ 8b | 6.4 (5.7, 7.1) | 0.6 (0.0, 1.2) | 4.9 (4.3, 5.5) | -0.3 (-0.8, 0.2) | 2.2 (1.8, 2.6) | 0.0 (-0.4, 0.3) | 1.8 (1.5, 2.1) | -0.6 (-1.0, -0.3) |
| Sample-specific cutoff of ≥ 8 – Population optimal cutoff of ≥ 8 | -0.8 (-1.7, 0.0) | 0.1 (-0.1, 0.4) | 0.2 (-0.3, 0.8) | -0.1 (-0.2, 0.1) | 0.1 (-0.2, 0.4) | 0.0 (-0.1, 0.1) | -0.1 (-0.4, 0.1) | 0.0 (-0.1, 0.1) |

Differences are presented as mean difference (95% confidence interval)

a Sample-specific optimal cutoff refers to the cutoff maximizing Youden’s J in each simulated sample

b the optimal cutoff in the full PHQ-9 IPDMA dataset is ≥ 8 (sensitivity = 80.4%, specificity = 82.0%)