Using individual participant data to evaluate the bias in cutoff identification and diagnostic accuracy estimates due to data–driven optimal cutoff selection

Parash Mani Bhandari

Department of Epidemiology, Biostatistics and Occupational Health,

McGill University, Montréal

May 2020

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science (Epidemiology)

Abstract

Background: Diagnostic accuracy studies often use small datasets to simultaneously select an "optimal" cutoff that maximizes accuracy and to estimate test accuracy. Clinicians are often encouraged to adopt the optimal cutoffs based on results from these small studies with small numbers of cases. No studies have examined, using actual participant data, the degree to which this practice generates inaccurate optimal cutoff values and biased estimates of sensitivity and specificity.

Objectives: To use actual participant data to evaluate for different sample sizes the degree to which data-driven optimal cutoff selection identifies cutoffs that diverge from the population optimal cutoff and biases accuracy estimates.

Methods: A dataset accrued for individual participant data meta-analyses (IPDMA) on the diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) to identify major depression was used. Eligible studies compared screening test scores with major depression diagnoses based on validated diagnostic interviews. To evaluate the bias due to data-driven optimal cutoff selection, from the IPDMA dataset, 1000 samples with 100, 200, 500 and 1000 participants were drawn and optimal cutoff and accuracy estimates in these samples were compared with those in the entire IPDMA dataset. The optimal cutoff was defined in the entire IPDMA dataset (population) and in each sample by maximizing the Youden's J (sensitivity + specificity - 1).

Results: The population optimal cutoff was ≥ 11. Optimal cutoffs ranged from ≥ 5 to ≥ 17 for samples of 100 and ≥ 8 to ≥ 13 for samples of 1000. On average, sensitivity was overestimated by 6.5% and specificity underestimated by 1.3% for samples of 100; sensitivity was overestimated by 1.4% and specificity underestimated by 1.0% for samples of 1000.

Conclusions: Data-driven optimal cutoff selection results in cutoffs that diverge substantially from the population optimal cutoff and produce biased accuracy estimates, especially with smaller samples. Clinicians should ideally use optimal cutoffs generated from well-conducted meta-analyses and should only adopt population-specific cutoffs if validated consistently in multiple primary studies. Primary diagnostic accuracy studies should preferably use a larger sample to estimate the optimal cutoff and estimate diagnostic accuracy.

Abrégé

Contexte: Les études de précision diagnostique utilisent souvent de petite ensembles de données pour sélectionner simultanément un seuil «optimal» qui maximise la précision et pour estimer la précision du test. Les cliniciens sont souvent encouragés à adopter les seuils optimaux en fonction des résultats de ces petites études qui contiennent un faible nombre de cas. Aucune étude n'a examiné, à l'aide des données réelles des participants, le degré auquel cette pratique génère des valeurs de seuils optimales inexactes et des estimations de sensibilité et de spécificité biaisées.

Objectifs: Utiliser des données réelles des participants afin d'évaluer, pour différentes tailles d'échantillon, le degré auquel la sélection du seuil optimal basée sur les données identifie les seuils qui s'écartent du seuil optimal de la population et biaisent les estimations de précision.

Méthodes: Un ensemble de données accumulées pour les méta-analyses de données individuelles des participants (IPDMA) sur la précision diagnostique de l'Édimbourg Postnatal Depression Scale (EPDS) pour identifier la dépression majeure a été utilisé. Les études admissibles ont comparé les résultats des tests de dépistage aux diagnostics de dépression majeure sur une base d'entretiens diagnostiques validés. Pour évaluer le biais dû à la sélection de la valeur-seuil optimale basée sur les données, à partir de l'ensemble de données IPDMA, 1000 échantillons avec 100, 200, 500 et 1000 participants ont été tirés et les estimations des valeurs-seuils et des précisions optimales dans ces échantillons ont été comparées à celles de l'ensemble de données IPDMA. Le seuil optimal a été défini dans l’ensemble des données IPDMA (population) et dans chaque échantillon en maximisant le J de Youden (sensibilité + spécificité - 1).

Résultas: Le seuil optimal de la population était ≥ 11. Les seuils optimaux variaient de ≥ 5 à ≥ 17 pour les échantillons de 100, et de ≥ 8 à ≥ 13 pour les échantillons de 1000. En moyenne, la

sensibilité a été surestimée de 6,5% et la spécificité a été sous-estimée de 1,3% pour des échantillons de 100; la sensibilité a été surestimée de 1,4% et la spécificité a été sous-estimée de 1,0% pour des échantillons de 1000.

Conclusions: La sélection de la valeur-seuil optimale basée sur les données entraîne des seuils qui divergent considérablement du seuil optimale de la population et produisent des estimations de précision biaisées, en particulier avec des échantillons plus petits. Les cliniciens devraient idéalement utiliser des seuils optimaux générés à partir de méta-analyses bien menées et ne devraient adopter des seuils spécifiques à la population que s'ils sont validés de manière cohérente dans plusieurs études primaires. Les études primaires de précision du diagnostic devraient de préférence utiliser un échantillon plus grand pour estimer le seuil optimal et estimer la précision du diagnostic.

Preface and author contributions

This is a manuscript-based thesis evaluating the extent to which data-driven optimal cutoff selection identifies cutoffs that diverge from the population optimal cutoff and biases accuracy estimates. It is presented in five chapters.

Chapter 1 and 2 introduce the thesis and provide a summary of the literature review. These two chapters were drafted by Parash Mani Bhandari and critically revised by Andrea Benedetti and Brett D. Thombs.

Chapter 3 presents the manuscript prepared for submission to the *British Medical Journal*. This manuscript utilized data from DEPRESSD EPDS Collaboration which consists of trainees, staffs, steering committee members, knowledge users and data contributors from all over the world. This manuscript was drafted by Parash Mani Bhandari with contribution from Brooke Levis, Brett D. Thombs and Andrea Benedetti. All authors and DEPRESSD EPDS Collaboration co-authors provided a critical review, approved the final manuscript and agreed to include it in this thesis. Exact contribution of all 89 authors and DEPRESSD EPDS Collaboration co-authors is provided in the manuscript in Chapter 3.

Chapter 4 provides a brief discussion of the thesis in relation to the evidence gap, research and clinical implications and anticipates the avenues for future research work. Chapter 5 presents the conclusions. These two chapters were drafted by Parash Mani Bhandari and critically revised by Andrea Benedetti and Brett D. Thombs. At the end, bibliography and all supplementary materials on methodology and results of the manuscript in Chapter 3 are provided.

Acknowledgements

I am extremely grateful to my supervisors Dr. Andrea Benedetti and Dr. Brett D. Thombs. Despite their academic commitments and hectic schedules, they always prioritized our meetings and made sure that I was able to meet all deadlines and graduate on time. This thesis is a product of their continuous guidance, support and motivation throughout my master's study. I truly appreciate having had this opportunity to learn and grow under their supervision.

Sincere thanks to all the DEPRESSD Project team members for the work support and post-work fun activities. I enjoyed being a part of this cheerful team. I am especially grateful to Dr. Brooke Levis who helped me at every phase of my master's study starting with selecting courses and navigating university deadlines to revising the research protocols and manuscripts. My sincere thanks to Julia Nordlund for her help in French translation of the thesis abstract.

I am grateful to all the collaborators who contributed their data and all study participants without whom this thesis would not have been possible. I am thankful to all the faculty and administrative staff at the Department of Epidemiology, Biostatistics and Occupational Health, McGill University. I also acknowledge the studentship I received from the Research Institute of the McGill University Health Centre.

On a more personal note, I am thankful for the unconditional support throughout my academic endeavours from my parents, Goma Bhandari and Ganga Dhar Bhandari. Finally, thanks to my wife, Dipika Neupane, for being with me at all times, for joining graduate school with me and for continuously motivating me in this amazing journey together. I could not have asked for more.

Table of contents

Abstract ..... **ii**

Abrégé ..... **iv**

Preface and author contributions ..... **vi**

Acknowledgements ..... **vii**

Table of contents ..... **viii**

List of figures ..... **ix**

List of appendices ..... **x**

List of abbreviations ..... **xi**

Chapter 1: Introduction **1**

Chapter 2: Literature review **2**

2.1 Screening 2

2.2 Sensitivity and specificity 2

2.3 Cutoff selection and reporting practice 3

2.4 Variability in optimal cutoffs selected from data-driven approach 4

2.5 Bias in accuracy estimates of optimal cutoffs selected from data-driven approach 4

2.6 Evidence gap 5

Chapter 3: Manuscript **7**

Chapter 4: Discussion **42**

4.1 Thesis summary and implications 42

4.2 Future research 43

Chapter 5: Conclusions **44**

Bibliography **45**

Appendices **47**

Data supplement for the manuscript in Chapter 3 47

List of figures

Figure 1: Flow diagram of study selection process 36

Figure 2: Variability in optimal cutoffs in 1,000 samples of size 100, 200, 500 and 1,000..... 37

Figure 3: Boxplots of accuracy estimates of the optimal cutoff in 1,000 samples of size 100, 200, 500 and 1,000 compared to the accuracy estimates of cutoff ≥ 11 in the population..... 38

Figure 4: Bias in accuracy estimates stratified across range of optimal cutoffs 39

Figure 5: Boxplots of accuracy estimates of the cutoff ≥ 11 in 1,000 samples of size 100, 200, 500 and 1,000 compared to the accuracy estimates of cutoff ≥ 11 in the population..... 40

List of appendices

eMethods1. Review of the cutoff reporting practice in the abstracts of studies on diagnostic accuracy of the EPDS 48

eMethods2. Search strategies 54

eFigure1. Distribution of optimal cutoffs for the 49 individual primary studies included in the IPDMA dataset 59

eTable1. Reasons for exclusion for all articles excluded at full-text level (N = 213) 60

eTable2. Characteristics of included 49 primary studies included in the IPDMA dataset 87

eTable3. Characteristics of eligible primary studies that did not provide data for the present study (N = 24) 97

eTable4. Frequencies of EPDS scores for cases and non-cases of major depression in the full IPDMA dataset 103

eTable5. Bias in accuracy estimates of sample-specific optimal cutoffs compared to the accuracy estimates from population optimal cutoff, stratified by the magnitude of optimal cutoff 104

List of abbreviations

CI	Confidence Interval
CIDI	Composite International Diagnostic Interview
CIHR	Canadian Institutes of Health Research
CIS-R	Clinical Interview Schedule – Revised
DEPRESSD	DEPRESsion Screening Data
DIGS	Diagnostic Interview for Genetic Studies
DIS	Diagnostic Interview Schedule
DSM	Diagnostic and Statistical Manual of Mental Disorders
EPDS	Edinburgh Postnatal Depression Scale
FRQ-S	Fonds de recherche du Québec - Santé
ICD	International Classification of Diseases
IPDMA	Individual Participant Data Meta-Analysis
MDD	Major Depressive Disorder
MDE	Major Depressive Episode
MINI	Mini International Neuropsychiatric Interview
ROC	Receiver Operating Characteristic
SCID	Structured Clinical Interview for DSM Disorders
STARD	Standards for Reporting of Diagnostic Accuracy Studies

Chapter 1: Introduction

Diagnostic accuracy studies frequently use small datasets to determine the cutoff that maximizes some criterion, such as combined sensitivity and specificity. In many studies, the same sample is also used to estimate the diagnostic accuracy of this "optimal" cutoff. Due to imprecision from small samples, different studies often identify a wide range of optimal cutoffs, even for similar participant groups.[1] In addition, accuracy estimates that are generated tend to be overly optimistic,[2-5] and may not be replicated in clinical practice.

To my knowledge, there is no study that has explored if the variability of optimal cutoffs from data-driven approach varies for samples with different sizes. The limited evidence available on variability of optimal cutoffs and bias in accuracy estimates due to data-driven approaches to select optimal cutoffs is based on simulated datasets generated using hypothetical distributions of test scores. However, the generalizability of results from simulated datasets depends on whether realistic models and parameter values were used in data generation, and simulating datasets from large actual participant dataset can be a superior approach especially when the population data distribution is not well defined.[6]

In this thesis, this evidence gap in literature is addressed by using actual participant data to explore how data-driven cutoff selection results in inaccurate optimal cutoff values and biases the accuracy estimates. Specifically, the thesis has two objectives: to estimate, across different sample sizes, using scores from the EPDS administered to women during pregnancy and postpartum to detect major depression, the degree to which study-level data-driven cutoff selection:

1) results in the selection of optimal cutoffs that differ from the population optimal cutoff derived from the full IPDMA "population" dataset, and
2) generates biased accuracy results compared to results from the population optimal cutoff derived from the full IPDMA "population" dataset.

Chapter 2: Literature review

2.1 Screening

Screening refers to the administration of tests or examinations to a large population to identify the potential presence of an illness or a medical condition. The objective of screening is to identify people early in a disease process prior to symptom onset in order to intervene prior to further progression. Whether screening should be recommended, however, depends on several factors including whether there is availability of treatment and support mechanism for people who screen positive and subsequently get diagnosis of an illness and whether the overall benefit outweighs the harms from screening.[7] This risk-benefit ratio depends on many factors including the performance of screening tests.

2.2 Sensitivity and specificity

The performance of screening tests is evaluated by several measures, though the two most commonly used measures are sensitivity and specificity.

Table 1: 2×2 table of test results vs. true disease status

		Disease status	
		Present	**Absent**
Test result	**Positive**	a	b
	Negative	c	d

Sensitivity = $\frac{a}{a+c}$

Specificity = $\frac{d}{b+d}$

Sensitivity of a test refers to the proportion of people with disease who will be correctly identified to have the disease. Specificity of a test refers to the proportion of people without disease who will be correctly identified to be disease-free. A test with poor sensitivity might result in missed cases and lost opportunities for treatment. On the other hand, a test with poor

specificity might lead to resources spent on expensive referrals and might expose people to invasive diagnostic tests and unnecessary treatments and their side effects.[8]

2.3 Cutoff selection and reporting practice

Many screening and diagnostic tests measure symptoms of participants in continuous or ordinal scale and provide a cumulative total score. For making decision about whether the score for participants can be considered a positive or a negative result, cutoff thresholds are needed. Establishing the cutoff threshold requires consideration of the financial and other costs of false negative and false positive results, and a comparative evaluation of the sensitivity and specificity estimates of all the relevant cutoffs. Receiver Operating Characteristic (ROC) curve is commonly used to summarize the sensitivity and specificity estimates of all potential cutoff thresholds. The goal is to identify a cutoff that has both high sensitivity and high specificity values. Selecting a higher cutoff, however, results in higher specificity but smaller sensitivity estimates and vice versa. Hence, a trade-off between maximizing sensitivity and specificity estimates must be made.

In practice, diagnostic accuracy studies often give equal weight to sensitivity and specificity and select the cutoff that maximizes Youden's J as the "optimal" cutoff. Youden's J is defined as: sensitivity + specificity – 1.[9] Its value ranges from 0 to 1. In a ROC curve, Youden's J optimal cutoff is the cutoff at the maximum vertical distance between the ROC curve and diagonal line.[10]

To explore cutoff selection and reporting practice in the abstracts of studies on the diagnostic test accuracy of Edinburgh Postnatal Depression Scale (EPDS), the most commonly used screening test for depression among pregnant and postpartum women,[11,12] a review was done. The objective was to explore how often accuracy was reported for one versus multiple cutoffs. In studies where accuracy was reported for one cutoff only, it was determined whether the cutoff was a data-driven optimal cutoff. It was found that of the 14 eligible studies, only one (7%) reported accuracy estimates for multiple cutoffs in the abstract, and 13 (93%)

reported accuracy for the study defined optimal cutoff only, whether or not that cutoff was standard (Appendices: eMethods1).

2.4 Variability in optimal cutoffs selected from data–driven approach

Many test accuracy studies use small samples to select the optimal cutoff. In the review described in Chapter 2.3, the sample size of the studies was found to range from 118 to 807 with an average sample of 320 participants (Appendices: eMethods1). However, in the evaluation of tests, samples are split into cases and non cases, and in many studies the number of cases is very small. Due to the use of relatively small sample sizes, optimal cutoffs identified in different test accuracy studies sometimes vary substantially, even when similar participant samples are used.[1] Authors of diagnostic test accuracy studies, however, often suggest that the data-driven optimal cutoffs that diverge from more standard cutoffs represent a unique optimal cutoff for the study's target population group, instead of being a result of small sample size or the data-driven approach used to select the optimal cutoff. For instance, in the review described in Chapter 2.3, no studies attributed a divergent optimal cutoff to a small sample size or to data-driven cutoff selection methods.

To my knowledge, this phenomenon of variability of data-driven optimal cutoffs in similar study samples has been explored by only one study. In that study, the authors explored the variability of optimal cutoffs with a simulation approach.[1] Using the published summary statistics, the authors drew 10,000 simulated samples, identified their optimal cutoff (obtained by maximizing Youden's J) and used the distribution of optimal cutoffs to quantify the variability. The authors found a high variability in optimal cutoffs identified and recommended that diagnostic accuracy studies should estimate and report variability along with the optimal cutoffs.

2.5 Bias in accuracy estimates of optimal cutoffs selected from data–driven approach

Accuracy estimates from optimal cutoffs selected using data-driven approach are likely to be biased. To my knowledge, only four studies have examined this phenomenon. The first study found that the total bias in sensitivity and specificity was maximum of 15% with a sample size

of 50 (N patients = 25, N controls = 25) and that the bias decreased as the sample size increased.[2] The second study compared this phenomenon with different prevalence rates and different number of patients and controls. The authors reported that the mean bias in sensitivity for a given specificity ranged from 1.5% to 5.6%, and, with a simulated sample of 250 and prevalence of 10%, which is approximately the prevalence rate for major depression in medical settings,[13,14] sensitivity was overestimated by 4.5%.[4] The third study, in a simulation with different number of patients and controls, different mean scores of patients and controls and different distributional assumptions, found that data-driven optimal cutoff overestimated the performance in simulated samples compared to that in the population from which the samples were drawn.[3] Finally, the fourth study also found that using data-driven optimal cutoff introduced bias in accuracy estimates and that an inverse relationship existed between the sample size and the bias in accuracy estimate.[5]

2.6 Evidence gap

The only study that evaluated variability in optimal cutoffs due to using a data-driven approach used summary statistics from actual studies but not the actual individual participant data to simulate samples and also did not evaluate the effect of sample size on the variability of optimal cutoffs. The four studies that evaluated bias in accuracy estimates from data-driven optimal cutoffs did not examine the degree to which bias in accuracy would occur when data-driven cutoffs are used to generate accuracy estimates for a specific screening or diagnostic test. More importantly, the authors used simulated datasets based on hypothetical distribution of test score rather than actual participant data.

The results from simulated datasets based on hypothetical distribution are as good as the accuracy of parametric model and the true parameters assumed for data generation. Especially when the population data distribution is not well defined, a much better approach for data generation is to use a large actual dataset from which data of desired size and frequency can be simulated (or resampled). Samples drawn this way mimic the actual dataset used for

resampling and if the actual dataset is big enough, estimates the true population parameters appropriately.[6]

Chapter 3: Manuscript

The following manuscript describes the research that was done to achieve the two thesis objectives listed in Chapter 1:

Bhandari PM, Levis B, Neupane D, Patten SB, Shrier I, Thombs BD, Benedetti A, and the DEPRESsion Screening Data (DEPRESSD) EPDS Collaboration. *Bias in cutoff identification and diagnostic accuracy estimates due to data–driven cutoff selection: simulation study using individual participant data from 49 studies on the diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS).*

The manuscript is prepared for submission to the *British Medical Journal* and is formatted accordingly.

Title:

Bias in cutoff identification and diagnostic accuracy estimates due to data-driven cutoff selection: simulation study using individual participant data from 49 studies on the diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS)

Authors:

Parash Mani Bhandari, Brooke Levis, Dipika Neupane, Scott B. Patten, Ian Shrier, Brett D. Thombs*, Andrea Benedetti*, and the DEPRESsion Screening Data (DEPRESSD) EPDS Collaboration

*Co-senior authors

DEPRESSD EPDS Collaboration:

Ying Sun, Chen He, Danielle B. Rice, Ankur Krishnan, Yin Wu, Marleine Azar, Tatiana A. Sanchez, Matthew J. Chiovitti, Nazanin Saadat, Kira E. Riehm, Mahrukh Imran, Zelalem Negeri, Jill T. Boruff, Pim Cuijpers, Simon Gilbody, John P.A. Ioannidis, Lorie A. Kloda, Roy C. Ziegelstein, Liane Comeau, Nicholas D. Mitchell, Marcello Tonelli, Simone N. Vigod, Franca Aceti, Rubén Alvarado, Cosme Alvarado-Esquivel, Muideen O. Bakare, Jacqueline Barnes, Amar D. Bavle, Cheryl Tatano Beck, Carola Bindt, Philip M. Boyce, Adomas Bunevicius, Tiago Castro e Couto, Linda H. Chaudron, Humberto Correa, Felipe Pinheiro de Figueiredo, Valsamma Eapen, Nicolas Favez, Ethel Felice, Michelle Fernandes, Bárbara F. C. Figueiredo, Jane R. W. Fisher, Lluïsa Garcia-Esteve, Lisa Giardinelli, Nadine Helle, Louise M. Howard, Dina Sami Khalifa, Jane Kohlhoff, Zoltán Kozinszky, Laima Kusminskas, Lorenzo Lelli, Angeliki A. Leonardou, Michael Maes, Valentina Meuti, Sandra Naki Radoš, Purificación Navarro García, Daisuke Nishi, Daniel Okitundu Luwa E-Andjafono, Susan J. Pawlby, Chantal Quispel, Emma Robertson-Blackmore, Tamsen J. Rochat, Heather J. Rowe, Deborah J. Sharp, Bonnie W. M. Siu, Alkistis Skalkidou, Alan Stein, Robert C. Stewart, Kuan-Pin Su, Inger Sundström-Poromaa, Meri Tadinac, S. Darius Tandon, Iva Tendais, Pavaani Thiagayson, Annamária Töreki, Anna Torres-Giménez, Thach D.

Tran, Kylee Trevillion, Katherine Turner, Johann M. Vega-Dienstmaier, Karen Wynter, Kimberly A. Yonkers

Affiliations:

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

Parash Mani Bhandari (masters student)

Dipika Neupane (masters student)

Ian Shrier (associate professor)

Brett D. Thombs (professor)

Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire, UK

Brooke Levis (postdoctoral fellow)

Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada

Scott B. Patten (professor)

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

Andrea Benedetti (associate professor)

Affiliations of DEPRESSD EPDS Collaboration Members:

Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada (Ying Sun, Chen He, Danielle B. Rice, Ankur Krishnan, Yin Wu, Marleine Azar, Tatiana A. Sanchez, Matthew J. Chiovitti, Nazanin Saadat, Kira E. Riehm, Mahrukh Imran, Zelalem Negeri); Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montréal, Québec, Canada (Jill T. Boruff); Department of Clinical, Neuro and Developmental

Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit, Amsterdam, the Netherlands (Pim Cuijpers); Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK (Simon Gilbody), Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA (John P.A. Ioannidis); Library, Concordia University, Montréal, Québec, Canada (Lorie A. Kloda); Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA (Roy C. Ziegelstein); International Union for Health Promotion and Health Education, École de santé publique de l'Université de Montréal, Montréal, Québec, Canada (Liane Comeau); Department of Psychiatry, University of Alberta, Edmonton, Alberta, Canada (Nicholas D. Mitchell); Department of Medicine, University of Calgary, Calgary, Alberta, Canada (Marcello Tonelli); Women's College Hospital and Research Institute, University of Toronto, Toronto, Ontario, Canada (Simone N. Vigod); Department of Neurology and Psychiatry, Sapienza University of Rome, Rome, Italy (Franca Aceti); School of Public Health, Faculty of Medicine, Universidad de Chile, Santiago, Chile (Rubén Alvarado); Laboratorio de Investigación Biomédica, Facultad de Medicina y Nutrición, Avenida Universidad, Dgo, Mexico (Cosme Alvarado-Esquivel); Child and Adolescent Unit, Federal Neuropsychiatric Hospital, Enugu, Nigeria (Muideen O. Bakare); Department of Psychological Sciences, Birkbeck, University of London, UK (Jacqueline Barnes); Department of Psychiatry, Rajarajeswari Medical College and Hospital, Bengaluru, Karnataka, India (Amar D. Bavle); University of Connecticut School of Nursing, Mansfield, Connecticut, USA (Cheryl Tatano Beck); Department of Child and Adolescent Psychiatry, University Medical Center Hamburg-Eppendorf, Germany (Carola Bindt); Discipline of Psychiatry, Westmead Clinical School, Sydney Medical School, University of Sydney, Sydney, Australia (Philip M. Boyce); Neuroscience Institute, Lithuanian University of Health Sciences, Kaunas, Lithuania (Adomas Bunevicius); Federal University of Uberlândia, Brazil (Tiago Castro e Couto); University of Rochester School of Medicine and Dentistry, Rochester, New York, USA (Linda H. Chaudron); Medicine Faculty - Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brazil (Humberto Correa); Department of Neurosciences and Behavior, Ribeirão Preto Medical School, Brazil

(Felipe Pinheiro de Figueiredo); University of New South Wales and Ingham Institute South West Sydney LHD, Australia (Valsamma Eapen); Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland (Nicolas Favez); Department of Psychiatry, Mount Carmel Hospital, Attard, Malta (Ethel Felice); Faculty of Medicine, Department of Paediatrics, University of Southampton, Southampton and Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK (Michelle Fernandes); School of Psychology, University of Minho, Portugal (Barbara Figueiredo); School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia (Jane R. W. Fisher); Perinatal Mental Health Unit CLINIC-BCN, Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain (Lluïsa Garcia-Esteve); Psychiatry Unit, Department of Health Sciences, University of Florence, Firenze, Italy (Lisa Giardinelli); Department of Child and Adolescent Psychiatry, University Medical Center Hamburg-Eppendorf, Germany (Nadine Helle); Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK (Louise M. Howard); Ahfad University for Women, Omdurman, Sudan (Dina Sami Khalifa); University of New South Wales, Kensington, Australia (Jane Kohlhoff); Department of Obstetrics and Gynaecology, Danderyd Hospital, Stockholm, Sweden (Zoltán Kozinszky); Private Practice, Hamburg, Germany (Laima Kusminskas); Psychiatry Unit, Department of Health Sciences, University of Florence, Firenze, Italy (Lorenzo Lelli); First Department of Psychiatry, Women's Mental Health Clinic, Athens University Medical School, Athens, Greece (Angeliki A. Leonardou); Department of Psychiatry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand (Michael Maes); Department of Neurology and Psychiatry, Sapienza University of Rome, Rome, Italy (Valentina Meuti); Department of Psychology, Catholic University of Croatia, Zagreb, Croatia (Sandra Naki Radoš); Perinatal Mental Health Unit CLINIC-BCN. Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain (Purificación Navarro García); Department of Mental Health, Graduate School of Medicine, The University of Tokyo, Japan (Daisuke Nishi); Unité de Neuropsychologie, Département de Neurologie, Centre Neuro-psycho-pathologique, Faculté de Médecine, Université de Kinshasa, République Démocratique du Congo (Daniel Okitundu Luwa E-Andjafono); Institute of Psychiatry, Psychology & Neuroscience, King's College London,

London, UK (Susan J. Pawlby); Department of Obstetrics and Gynaecology, Albert Schweitzer Ziekenhuis, Dordrecht, the Netherlands (Chantal Quispel); Halifax Health, Graduate Medical Education, Daytona Beach, FL. USA (Emma Robertson-Blackmore); MRC/Developmental Pathways to Health Research Unit, School of Clinical Medicine, University of Witwatersrand, South Africa (Tamsen J. Rochat); School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia (Heather J. Rowe); Centre for Academic Primary Care, Bristol Medical School, University of Bristol, UK (Deborah J. Sharp); Department of Psychiatry, Castle Peak Hospital, Hong Kong SAR, China (Bonnie W. M. Siu); Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden (Alkistis Skalkidou); University of Oxford, Oxford, UK (Alan Stein); Department of Mental Health, College of Medicine, University of Malawi, Malawi (Robert C. Stewart); College of Medicine, China Medical University, Taichung, Taiwan (Kuan-Pin Su); Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden (Inger Sundström-Poromaa); Department of Psychology, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia (Meri Tadinac); Northwestern University Feinberg School of Medicine, Chicago, IL, USA (S. Darius Tandon); School of Psychology, University of Minho, Portugal (Iva Tendais); Institute of Mental Health, Singapore (Pavaani Thiagayson); Department of Emergency, University of Szeged, Hungary (Annamária Töreki); Perinatal Mental Health Unit CLINIC-BCN. Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain (Anna Torres-Giménez); School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia (Thach D. Tran); Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK (Kylee Trevillion); Epilepsy Center-Child Neuropsychiatry Unit, ASST Santi Paolo Carlo, San Paolo Hospital, Milan, Italy (Katherine Turner); Facultad de Medicina Alberto Hurtado, Universidad Peruana Cayetano Heredia. Lima, Perú (Johann M. Vega-Dienstmaier); School of Nursing & Midwifery, Deakin University, Melbourne, Australia (Karen Wynter); and Department of Psychiatry, Yale School of Medicine, New Haven, Connecticut, USA (Kimberly A. Yonkers).

Corresponding authors

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

Brett D. Thombs, PhD; Jewish General Hospital; 4333 Cote Ste. Catherine Road; Montreal, Quebec, Canada H3T 1E4; Tel (514) 340-8222 ext. 25112; E-mail: brett.thombs@mcgill.ca

Word count: 3622

ABSTRACT

Objective: Studies of test accuracy often use small datasets to simultaneously select an "optimal" cutoff that maximizes test accuracy and generate the accuracy estimates. Our objective was to use actual participant data across a range of sample sizes to evaluate the degree to which data-driven optimal cutoff selection with the Edinburgh Postnatal Depression Scale (EPDS) to detect major depression results in (1) selection of optimal cutoffs that diverge from the population optimal cutoff and (2) biased accuracy estimates.

Design: Simulation study in which 1,000 samples each of sample sizes 100, 200, 500 and 1,000 were randomly drawn with replacement from a "population" sample of actual participant data.

Data sources: Dataset accrued for an individual participant data meta-analysis on EPDS accuracy to detect major depression in pregnancy and postpartum.

Eligibility criteria for selecting studies: Eligible studies compared EPDS scores with major depression classification based on validated diagnostic interviews. For the study population and for each simulated sample, an optimal cutoff was selected by maximizing Youden's J (sensitivity + specificity – 1). Accuracy estimates for the optimal cutoff in each simulated sample were compared to the accuracy estimate in the population.

Results: The population included 13,255 participants from 49 primary studies. Population optimal cutoff was ≥ 11. Optimal cutoffs in individual samples ranged from ≥ 5 to ≥ 17 for N = 100, ≥ 6 to ≥ 16 for N = 200, ≥ 6 to ≥ 14 for N = 500, and ≥ 8 to ≥ 13 for N = 1,000. The percentage of studies that identified the true population optimal cutoff was 30% for N = 100, 35% for N = 200, 53% for N = 500, and 71% for N = 1,000. Mean overestimation of sensitivity and mean underestimation of specificity were 6.5% and -1.3% for N = 100, 4.2% and -1.1% for N = 200, 1.8% and -1.0% for N = 500, and 1.4% and -1.0% for N = 1000.

Conclusions: Data-driven optimal cutoff selection resulted in cutoffs that diverged substantially from the population optimal cutoff and produced biased accuracy estimates, especially with smaller samples. Clinicians and researchers should use cutoffs identified in large meta-analyses or, alternatively, identified consistently across multiple primary studies.

INTRODUCTION

Many screening and diagnostic tests measure symptoms on a continuous or ordinal scale and use a cutoff threshold to distinguish between positive and negative results compared to a diagnostic standard. Data-driven approaches are often used in studies with relatively small samples to both select an "optimal" cutoff threshold to maximize some criterion, such as combined sensitivity and specificity, and to draw conclusions about the accuracy of the test.[1] As a result, optimal cutoffs identified in different studies sometimes vary substantially, even when similar participant samples are used. In addition, accuracy estimates tend to be overly optimistic because the threshold is set to maximize accuracy in the study sample, even though it may not maximize accuracy in other samples or in the target population and may not represent what would occur in clinical practice.

We know of only four studies that have investigated the degree to which data-driven selection of cutoff scores may influence diagnostic accuracy estimates.[2-5] These studies each reported that data-driven cutoff selection produces overly optimistic estimates, particularly in small sample sizes. However, all of these studies used simulated datasets based on hypothetical test score distributions rather than actual participant data. Thus, how widely data-driven optimal cutoffs diverge from population-based optimal cutoffs and how biased estimates of diagnostic accuracy are is not known for any specific screening or diagnostic test.

Depression screening in pregnancy and the postpartum period is often recommended,[6, 7] and the Edinburgh Postnatal Depression Scale (EPDS) is the most commonly used screening tool.[8, 9] A recent individual participant data meta-analysis (IPDMA) found that a cutoff of ≥ 11 maximized combined sensitivity and specificity.[10] Primary studies that assess the accuracy of the EPDS often set cutoffs and attempt to estimate screening accuracy based on data-driven methods. In many of these studies, the abstract, which may be the only part of the article that is read,[11, 12] reports only results from a single data-driven optimal cutoff. We reviewed 14 recently published primary studies of EPDS accuracy (range of number of participants = 118 to 807; mean = 320) and found that only one of them reported accuracy results from more than one cutoff in the abstract. The remaining 13 (93%) reported only results from the single best-

performing cutoff; 11 of the 13 (85%) maximized Youden's J (sensitivity + specificity – 1) to select the best-performing cutoff. Cutoffs identified as optimal in the 14 studies ranged from ≥ 8 to ≥ 13. In some studies, authors asserted that the data-driven optimal cutoffs, which diverged from more standard cutoffs, represented a unique optimal cutoff for the study's target population group. No studies attributed a divergent optimal cutoff to a small sample size or to data-driven cutoff selection methods (see Appendix: eMethods1).

The objectives of the present study were to estimate, across different sample sizes, using scores from the EPDS administered to women during pregnancy and postpartum to detect major depression, the degree to which study-level data-driven cutoff selection: (1) results in the selection of optimal cutoffs that differ from the population optimal cutoff derived from the full IPDMA "population" dataset and (2) generates biased accuracy results compared to results from the population optimal cutoff.

METHODS

We used data accrued for an IPDMA on the accuracy of the EPDS for depression screening[10] to form a study population from which to simulate the sampling of individual primary studies of different sample sizes. A protocol for the present study was uploaded to the Open Science Framework repository prior to initiating the study (https://osf.io/qnvzp/).

Study eligibility

Datasets from articles in any language were eligible for inclusion in the main IPDMA and the present study if: (1) they included EPDS scores; (2) they included diagnostic classification for current Major Depressive Disorder (MDD) or Major Depressive Episode (MDE) using Diagnostic and Statistical Manual of Mental Disorders (DSM)[13-15] or International Classification of Diseases (ICD)[16] criteria based on a validated semi-structured or fully structured interview; (3) the interview and EPDS were administered within two weeks of each other because DSM and ICD major depression diagnostic criteria specify that symptoms must have been present in the last two weeks; (4) participants were women ≥ 18 years who were pregnant or had given birth in the previous 12 months; and (5) participants were not recruited from psychiatric settings or because they were suspected of having depression, since screening is done to identify

previously unrecognized cases.[17] Datasets where not all participants were eligible were included if primary data allowed selection of eligible participants.

Database searches and study selection

A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations and PsycINFO via OvidSP, and Web of Science via ISI Web of Knowledge from inception to June 10, 2016, using a peer-reviewed[18] search strategy (Appendix: eMethods2). We additionally reviewed reference lists of relevant reviews and queried contributing authors about non-published studies, including studies in progress or submitted for peer review. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada) for processing review results.

Two investigators independently reviewed titles and abstracts for eligibility. If either deemed a study potentially eligible, independent full-text review was done by two investigators with disagreements resolved by consensus, consulting a third investigator when necessary. Translators were consulted for languages other than those in which team members were fluent.

Data contribution, extraction, and synthesis

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

Individual participant data were converted to a standard format and synthesized into a single dataset. We compared published participant characteristics and diagnostic accuracy results with results from raw datasets and resolved any discrepancies in consultation with the original investigators. For the present study, if the study collected data at multiple time points, we included data from only the time point with the most participants. If there was a tie, we selected the time point with the largest number of major depression cases. For defining major depression, we considered MDD or MDE based on the DSM or ICD. If more than one was

reported, we prioritized MDE over MDD (since screening would attempt to detect depressive episodes and further interview would determine whether the episode was related to MDD, bipolar disorder, or persistent depressive disorder), and we prioritized DSM over ICD.

Simulation of study samples and statistical analyses

Unlike many other depression screening tools, the EPDS does not have a clearly recognized standard cutoff for depression screening. The original validation study which included 84 participants and 24 cases of definite or probable major depression based on Research Diagnostic Criteria suggested that cutoffs of ≥ 10 or ≥ 13 could be used.[19] However, many studies report using different cutoffs between ≥ 10 and ≥ 13 to identify major depression,[20, 21] with ≥ 13 being the most commonly used cutoff.[21] A recent IPDMA using a slightly larger version of the dataset in the present study found that a cutoff of ≥ 11 maximized Youden's J (sensitivity + specificity -1) overall and for subgroups.[10]

For the present study, we used our IPDMA dataset to represent a hypothetical "population" of women, and we defined population sensitivity and specificity values for EPDS cutoffs to be those estimated in this population. To do this, we analyzed the IPDMA dataset, ignoring sampling weights as well as study-level clustering of observations. We ignored sampling weights and clustering in order to have a defined population from which we could draw samples that represented simulated primary studies and to be able to use the same analytical approach when analyzing the population data and the simulated primary study data. As a result, we generated accuracy estimates that differed slightly from those reported in the full EPDS IPDMA, which used sampling weights and study-level clustering and a slightly larger sample.[10] We verified that a cutoff of ≥ 11 maximized Youden's J for the unweighted population.

From the overall population IPDMA dataset, we sampled with replacement to generate 1,000 random samples each with N = 100, N = 200, N = 500, and N = 1,000 participants. For each sample, we defined the sample-specific optimal cutoff as the cutoff that maximized Youden's J in the sample. If there was a tie in maximum Youden's J between multiple cutoffs, we selected the higher cutoff as the sample-specific optimal cutoff. For each sample size,

across the 1,000 samples, we (1) graphically illustrated the variability in sample-specific optimal cutoffs based on Youden's J and the variability in accuracy of the sample-specific optimal cutoffs; (2) estimated the mean difference and associated 95% confidence interval (CI) between sensitivity and specificity based on sample-specific optimal cutoffs versus the population sensitivity and specificity based on the population optimal cutoff of ≥ 11, and (3) estimated the mean difference and 95% CI for sensitivity and specificity based on a cutoff of ≥ 11 in each sample versus the population sensitivity and specificity also based on a cutoff of ≥ 11. CIs for the variability in optimal cutoffs and the unweighted accuracy estimates were computed using a one sample proportion test with continuity correction. For all analyses, sensitivity and specificity were estimated using crude 2 x 2 table counts. In additional analyses, we stratified results by the optimal cutoff value identified in each sample.

Patient and public involvement

Patients and the public were not involved in the design, conduct or reporting of this study.

Deviations from protocol

We initially specified that we would also compare accuracy of the optimal cutoff in each sample with that of cutoff ≥ 13, which is the cutoff most commonly used in practice.[18-20] We subsequently determined that a population optimal cutoff of ≥ 11 maximizes Youden's J in our IPDMA "population", which was established in the main EPDS IPDMA dataset[10] and confirmed in the present study. Thus, we used a cutoff of ≥ 11 only and not ≥ 13.

RESULTS

Study search results and inclusion of primary data

Of 3,417 unique titles and abstracts identified from the database search, 3,097 were excluded after title and abstract review and 213 were excluded after full-text review, leaving 107 eligible articles with data from 72 unique participant samples, of which 48 (66.7%) contributed datasets (Figure 1). Reasons for exclusion for each article excluded at the full-text level are provided in Appendix: eTable1. In addition, authors of included studies contributed data from one unpublished study, which was subsequently published, for a total of 49

datasets. Characteristics of eligible included primary studies are provided in Appendix: eTable2, and characteristics of eligible primary studies that did not contribute data are shown in Appendix: eTable3. In total, 13,255 participants (1,625 major depression cases [12.3%]) were included. Frequencies of EPDS scores for cases and non-cases in the full database are shown in Appendix: eTable4.

The sample size of the 49 included primary studies ranged from 40 to 2,634 with a mean of 271 (median = 190). The mean number of cases of major depression was 34 (median = 25), and 20 studies included < 20 cases of major depression. As shown in Appendix: eFigure1, study-specific optimal cutoffs that maximized Youden's J ranged from ≥ 5 to ≥ 19. For the "population" of 13,255 participants and using a cutoff of ≥ 11, the unweighted sensitivity and specificity were 78.7% (95% CI: 76.6% to 80.7%) and 83.4% (95% CI: 82.7% to 84.0%).

Variability of sample–specific optimal cutoffs in simulated samples

Figure 2 shows the variability of data-driven sample-specific optimal cutoffs from 1,000 samples each of 100, 200, 500 and 1,000 participants.

Optimal cutoffs in individual samples ranged from ≥ 5 to ≥ 17 for N = 100, ≥ 6 to ≥ 16 for N = 200, ≥ 6 to ≥ 14 for N = 500, and ≥ 8 to ≥ 13 for N = 1,000. The percentage of studies that identified the true population optimal cutoff of ≥ 11 was 30.3% (95% CI: 27.5% to 33.3%) for N = 100, 34.7% (95% CI: 31.8% to 37.8%) for N = 200, 53.0% (95% CI: 49.9% to 56.1%) for N = 500, and 70.5% (95% CI: 67.6% to 73.3%) for N = 1,000.

Bias from data–driven cutoff selection in simulated samples

As shown in Figure 3 and Table 1, based on the overall mean across 1,000 samples, sensitivity based on sample-specific optimal cutoffs was overestimated compared to the sensitivity in the population based on the population optimal cutoff by 6.5% (95% CI: 5.8% to 7.2%) for N = 100, 4.2% (95% CI: 3.6% to 4.7%) for N = 200, 1.8% (95% CI: 1.4% to 2.1%) for N = 500 and 1.4% (95% CI: 1.1% to 1.6%) for N = 1,000. In turn, specificity was underestimated by 1.3% (95% CI: -1.9% to -0.7%) for N = 100, 1.1% (95% CI: -1.6% to -0.7%) for N = 200, 1.0% (95% CI: -1.3% to -0.7%) for N = 500 and 1.0% (95% CI: -1.2% to -0.8%) for N = 1,000.

Figure 4 and Appendix: eTable5 show that the direction and magnitude of bias in sensitivity and specificity estimates depended on the optimal cutoff identified in each sample. For instance, with a sample size of 100, in samples with sample-specific optimal cutoff ≥ 5 to ≥ 8, sensitivity was overestimated by 16.0% (95% CI: 14.8% to 17.2%), and specificity was underestimated by 19.6% (95% CI: -20.8% to -18.3%). For samples with sample-specific optimal cutoffs of ≥ 14 to ≥ 17, sensitivity was underestimated by 6.3% (95% CI: -8.9% to -3.7%), and specificity was overestimated by 10.7% (95% CI: 10.2% to 11.2%).

As shown in Figure 5, when sensitivity and specificity were calculated for cutoff ≥ 11 in each sample, the mean sensitivity and specificity were close to that of the population values. Variability patterns were similar to those observed using sample-specific optimal cutoffs. See also Table 1.

DISCUSSION

We used actual participant EPDS data from over 13,000 participants from 49 primary studies and simulated 1,000 primary studies each for sample sizes of 100, 200, 500, and 1,000. There were two main findings. First, with very small sample sizes (N = 100), study-specific optimal cutoffs ranged from ≥ 5 to ≥ 17 when the actual population optimal cutoff was ≥ 11. Even with samples of N = 1,000, optimal cutoffs ranged from ≥ 8 to ≥ 13. Second, for samples of N = 100, mean overestimation of sensitivity was 6.5% whereas mean underestimation of specificity was 1.3%. For larger samples (N = 1,000), sensitivity was overestimated, on average, by 1.4% and specificity underestimated by 1.0%. The degree and direction of sample-specific deviance from population-level estimates depended on the identified sample-specific optimal cutoff. For N = 100, for example, individual studies that identified optimal cutoffs from ≥ 5 to ≥ 8 overestimated sensitivity by an average of 16.0%; studies that identified high optimal cutoffs (≥ 14 to ≥ 17), on the other hand, underestimated sensitivity by 6.3%.

Findings in context

The degree of variability we identified in sample-specific optimal cutoffs, especially with smaller sample sizes, is concerning, because most diagnostic accuracy studies of depression

screening tools are conducted in small samples. Among the 49 studies included in the present IPDMA dataset, 26 (53.1%) had sample size of < 200, 19 (38.8%) had sample size of 200 to 500, 3 (6.1%) had sample size of 501 to 1000 and only one (2.0%) had sample size > 1000. A previous study examined sample sizes and the presence of sample size calculations in 89 studies of depression screening tool accuracy, not limited to the EPDS, and found that the median sample size was 224; 38 (42.7%) had sample size of < 200, 33 (37.1%) had sample size of 200 to 500, 11 (12.3%) had sample size of 501 to 1000 and 7 (7.9%) had sample size of > 1000.[22] Based on our findings, we would expect that, overall, many studies of test accuracy of depression screening tools likely overestimate sensitivity with only minor losses in specificity. A larger bias in sensitivity estimates as compared to the bias in specificity estimates is due to the presence of fewer number of participants with major depression as compared to those without major depression. Due to this, optimal cutoff selection in some samples resulted in substantial gains in sensitivity with relatively small compensation in specificity, particularly in small samples. As shown in the present study, however, mean differences do not capture what may occur in any given study, and depending on the specific sample, sensitivity may be overestimated or underestimated, sometimes substantially.

Research and clinical implications

The Standards for Reporting of Diagnostic Accuracy Studies (STARD) reporting guideline recommends *a priori* sample size estimation for the desired precision level in accuracy estimates.[23] Results from our study show that setting sample size targets pre-study should also consider variability in the optimal cutoff that may be identified and not just variability in accuracy estimates. A previous study that examined sample sizes in 89 studies of depression screening tool accuracy found that only 3 reported *a priori* sample size calculations, and none of these specifically considered the issue of identifying an optimal cutoff and estimating accuracy in the same participant sample.[22] Authors of primary studies on depression screening tool accuracy could potentially use bootstrapping, parametric or non-parametric methods to estimate confidence intervals for uncertainty around the optimal cutoff.[24, 25] They could also employ internal validation methods like cross-sampling, sample-splitting and bootstrapping to

statistically adjust for the bias in accuracy estimates from data-driven optimal cutoff selection.[26] These methods, however, have not been demonstrated or tested in the context of mental health screening. Indeed, the most robust approach for identifying optimal cutoffs and generating accurate estimates of screening or diagnostic accuracy is through pooling large numbers of well-conducted primary studies and participants via meta-analysis, preferably IPDMA, which can ensure that all cutoffs are available for examination for all participants.[27, 28] To facilitate this, researchers should report accuracy data from primary accuracy studies for all possible cutoffs in 2×2 form, at least in appendices, to facilitate subsequent synthesis and to avoid selective cutoff reporting bias.[29]

Administering diagnostic interviews to large samples can be resource-intensive. A less resource-intensive alternative might be to administer a diagnostic interview to participants with positive screens but to only a proportion of those with negative screens. Weighting can then be used to generate accuracy estimates to reflect the actual proportion of participants in the study sample.[30] With sufficient numbers of participants and cases, this can be done with relatively minimal loss of precision compared to conducting interviews with all study participants.[31]

As shown in our review of recent studies of the EPDS (Appendix: eMethods1), authors of diagnostic accuracy studies that identify study-specific optimal cutoffs that depart from standard cutoffs often conclude that this is evidence for the need to use different cutoffs in different populations. While this is possible, our full IPDMA of the screening accuracy of the EPDS did not find evidence for differential accuracy by subgroups.[10] Results from the present study suggest that variability in identified optimal cutoffs most often results from chance variability alone and may not reflect the existence of any group differences. Hence, authors of individual studies should avoid recommending specific cutoffs for specific populations unless the studies use large samples or the findings are replicated consistently across multiple studies. Researchers who conduct trials of screening and health care providers who wish to use the EPDS in practice should select cutoffs from large, well-conducted meta-analyses. They

should be aware that primary study-specific results are often not accurate and should not guide clinical practice.

Strengths and limitations of study

Strengths of this study include the use of actual participant data instead of simulated data and hypothetical distributional assumptions, the large population sample size from which we were able to sample, and the ability to include results from studies that collected eligible primary data even if accuracy results were not published, which had not been done previously. There are also limitations to consider. We were unable to include primary data from 24 of 72 (33.3%) identified unique eligible participant samples. Nonetheless, we believe that the IPDMA database, with over 13,000 participants, was sufficiently large to simulate sampling in individual studies. Our results on the bias in sensitivity and specificity estimates is based on the analysis of a large dataset on depression screening accuracy of the EPDS, and the results may be different for a different test or a different study sample. Also, we estimated optimal cutoff by maximizing Youden's J, which gives equal weight to the estimates of sensitivity and specificity, and a different method of optimizing cutoff, for instance maximizing sensitivity alone, may change the observed bias in accuracy estimates. Despite these limitations, this study is the first to use actual participant data and provide evidence of the bias in accuracy estimates from using data-driven optimal cutoffs.

Conclusions and implications

Using the same participant sample to both identify an optimal cutoff and estimate screening or diagnostic accuracy often results in optimal cutoffs that differ from the population optimal cutoff and in accuracy estimates that are overly optimistic; these biases can be substantial when sample sizes are 100-500, which is common in published diagnostic accuracy studies of depression screening tools. Researchers who conduct primary studies of diagnostic accuracy should calculate sample sizes *a priori* and describe related limitations; they should avoid recommending cutoffs for population subgroups when sample sizes are not sufficiently large to support this; and they should report results completely so that they are easily synthesized in meta-analyses. Clinicians should be aware that conclusions about cutoffs

and accuracy from single studies may not reflect what will occur in practice and should preferably base decisions on a cutoff derived from well-conducted meta-analyses.

Contributors: PMBhandari, BL, DNeupane, JTB, PC, SG, JPAI, LAK, SBP, IS, RCZ, LC, NDM, MTonelli, SNV, BDT and ABenedetti were responsible for the study conception and design. JTB and LAK designed and conducted database searches to identify eligible studies. FA, RA, CAE, MOB, JB, ADB, CTB, CB, PMBoyce, ABunevicius, TCeC, LHC, HC, FPF, VE, NF, EF, MF, BF, JRWF, LGE, LG, NH, LMH, DSK, JK, ZK, LK, LL, AAL, MM, VM, SNR, PNG, DNishi, DOLEA, SJP, CQ, ERB, TJR, HJR, DJS, BWMS, ASkalkidou, AStein, RCS, KPS, ISP, MTadinac, SDT, IT, PT, AT, ATG, TDT, KTrevillion, KTurner, JMVD, KW and KAY contributed primary datasets that were included in this study. PMBhandari, BL, DNeupane, YS, CH, DBR, AK, YW, MA, TAS, MJC, NS, KER, MI and ZN contributed to data extraction and coding for the meta-analysis. PMBhandari, BL, BDT and ABenedetti contributed to the data analysis and interpretation. PMBhandari, BL, BDT and ABenedetti, contributed to drafting the manuscript. All authors provided a critical review and approved the final manuscript. BDT and ABenedetti are the guarantors; 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.

being asked to pay an open access fee—see http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/copyright-open-access-and-permission-reuse). The terms of such open access shall be governed by a Creative Commons licence—details as to which Creative Commons licence will apply to the research article are set out in our worldwide licence referred to above.

Funding: This study was funded by the Canadian Institutes of Health Research (CIHR, KRS-140994). Mr. Bhandari was supported by a studentship from the Research Institute of the McGill University Health Centre. Ms. Levis was supported by a CIHR Frederick Banting and Charles Best Canada Graduate Scholarship doctoral award and a Fonds de recherche du Québec - Santé (FRQ-S) Postdoctoral Training Award. Ms. Neupane was supported by G.R. Caverhill Fellowship from the Faculty of Medicine, McGill University. Ms. Rice was supported by a Vanier Canada Graduate Scholarship. Dr. Wu was supported by an Utting Postdoctoral Fellowship from the Jewish General Hospital, Montreal, Quebec, Canada and a FRQ-S Postdoctoral Training Award. Ms. Azar was supported by a FRQ-S Masters Training Award. The primary study by Alvarado et al. was supported by the Ministry of Health of Chile. The primary study by Barnes et al. was supported by a grant from the Health Foundation (1665/608). The primary study by Beck et al. was supported by the Patrick and Catherine Weldon Donaghue Medical Research Foundation and the University of Connecticut Research Foundation. The primary study by Helle et al. was supported by the Werner Otto Foundation, the Kroschke Foundation, and the Feindt Foundation. Prof. Robertas Bunevicius, MD, PhD (1958-2016) was Principal Investigator of the primary study by Bunevicius et al, but passed away and was unable to participate in this project. The primary study by Couto et al. was supported by the National Counsel of Technological and Scientific Development (CNPq) (Grant no. 444254/2014-5) and the Minas Gerais State Research Foundation (FAPEMIG) (Grant no. APQ-01954-14). The primary study by Chaudron et al. was supported by a grant from the National Institute of Mental Health (grant K23 MH64476). The primary study by Figueira et al. was supported by the Brazilian Ministry of Health and by the National Counsel of Technological and Scientific Development (CNPq) (Grant

no. 403433/2004-5). The primary study by de Figueiredo et al. was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo. The primary study by Tissot et al. was supported by the Swiss National Science Foundation (grant 32003B 125493). The primary study by Fernandes et al. was supported by grants from the Child: Care Health and Development Trust and the Department of Psychiatry, University of Oxford, Oxford, UK, and by the Ashok Ranganathan Bursary from Exeter College, University of Oxford. Dr. Fernandes is supported by a University of Southampton National Institute for Health Research (NIHR) academic clinical fellowship in Paediatrics. The primary study by Tendais et al. was supported under the project POCI/SAU-ESP/56397/2004 by the Operational Program Science and Innovation 2010 (POCI 2010) of the Community Support Board III and by the European Community Fund FEDER. The primary study by Fisher et al. was supported by a grant under the Invest to Grow Scheme from the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs. The primary study by Garcia-Esteve et al. was supported by grant 7/98 from the Ministerio de Trabajo y Asuntos Sociales, Women's Institute, Spain. The primary study by Howard et al. was supported by the NIHR under its Programme Grants for Applied Research Programme (Grant Reference Numbers RP-PG-1210-12002 and RP-DG-1108-10012) and by the South London Clinical Research Network. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The primary study by Phillips et al. was supported by a scholarship from the National Health and Medical and Research Council (NHMRC). The primary study by Roomruangwong et al. was supported by the Ratchadaphiseksomphot Endowment Fund 2013 of Chulalongkorn University (CU-56-457-HR). The primary study by Naki Radoš et al. was supported by the Croatian Ministry of Science, Education, and Sports (134-0000000-2421). The primary study by Navarro et al. was supported by grant 13/00 from the Ministry of Work and Social Affairs, Institute of Women, Spain. The primary study by Usuda et al. was supported by Grant-in-Aid for Young Scientists (A) from the Japan Society for the Promotion of Science (primary investigator: Daisuke Nishi, MD, PhD), and by an Intramural Research Grant for Neurological and Psychiatric Disorders from the National Center of Neurology and Psychiatry, Japan. The primary study by Pawlby et al. was supported

by a Medical Research Council UK Project Grant (number G89292999N). The primary study by Quispel et al. was supported by Stichting Achmea Gezondheid (grant number z-282). Dr. Robertson-Blackmore was supported by a Young Investigator Award from the Brain and Behavior Research Foundation and NIMH grant K23MH080290. The primary study by Rochat et al. was supported by grants from University of Oxford (HQ5035), the Tuixen Foundation (9940), and the Wellcome Trust (082384/Z/07/Z and 071571), and the American Psychological Association. Dr. Rochat receives salary support from a Wellcome Trust Intermediate Fellowship (211374/Z/18/Z). The primary study by Rowe et al. was supported by the diamond Consortium, beyondblue Victorian Centre of Excellence in Depression and Related Disorders. The primary study by Comasco et al. was supported by funds from the Swedish Research Council (VR: 521-2013-2339, VR:523-2014-2342), the Swedish Council for Working Life and Social Research (FAS: 2011-0627), the Marta Lundqvist Foundation (2013, 2014), and the Swedish Society of Medicine (SLS-331991). The primary study by Prenoveau et al. was supported by The Wellcome Trust (grant number 071571). The primary study by Stewart et al. was supported by Professor Francis Creed's Journal of Psychosomatic Research Editorship fund (BA00457) administered through University of Manchester. The primary study by Su et al. was supported by grants from the Department of Health (DOH94F044 and DOH95F022) and the China Medical University and Hospital (CMU94-105, DMR-92-92 and DMR94-46). The primary study by Tandon et al. was supported by the Thomas Wilson Sanitarium. The primary study by Tran et al. was supported by the Myer Foundation who funded the study under its Beyond Australia scheme. Dr. Tran was supported by an early career fellowship from the Australian National Health and Medical Research Council. The primary study by Vega-Dienstmaier et al. was supported by Tejada Family Foundation, Inc, and Peruvian-American Endowment, Inc. The primary study by Yonkers et al. was supported by a National Institute of Child Health and Human Development grant (5 R01HD045735). Drs. Benedetti and Thombs were supported by FRQ-S Researcher Salary Awards. No other authors reported funding for primary studies or for their work on the present study.

Declaration of competing interests: All authors have completed the ICMJE uniform disclosure form and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years with the following exceptions: Dr. Patten reports grants from Hotchkiss Brain Institute/Pfizer Competition, outside the submitted work. Dr. Tonelli declares that he has received a grant from Merck Canada, outside the submitted work. Dr. Vigod declares that she receives royalties from UpToDate, outside the submitted work. Dr. Beck declares that she receives royalties for her Postpartum Depression Screening Scale published by Western Psychological Services. Dr. Boyce declares that he receives grants and personal fees from Servier, grants from Lundbeck, and personal fees from AstraZeneca, all outside the submitted work. Dr. Howard declares that she has received personal fees from NICE Scientific Advice, outside the submitted work. Dr. Sundström-Poromaa declares that she has served on advisory boards and acted as invited speaker at scientific meetings for MSD, Novo Nordisk, Bayer Health Care, and Lundbeck A/S. Dr. Yonkers declares that she receives royalties from UpToDate, outside 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.

Ethics statement: As this study involved only analysis of previously collected de-identified data and because all included studies were required to have obtained ethics approval and informed consent, the Research Ethics Committee of the Jewish General Hospital determined that ethics approval was not required.

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

Data sharing: Requests to access data should be made to the corresponding authors.

What is already known on this topic:

- Previous studies have reported that data-driven optimal cutoff selection produces overly optimistic diagnostic accuracy estimates that cannot be replicated in clinical practice.
- All previous studies, however, have evaluated bias in simulated datasets with hypothetical distributions of test scores, not from actual participant data.

What this study adds:

- Data-driven selection of optimal cutoffs often results in selection of cutoffs that differ substantially from the population optimal cutoff, as well as, on average, in substantial overestimation of sensitivity with minimal underestimation of specificity, particularly with small samples.
- Clinicians should not use population-specific cutoffs identified in single studies unless those cutoffs and accuracy estimates are verified across multiple well-conducted studies or in a meta-analysis.

REFERENCES

1. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Med* 2016;26:297-307.
2. Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin Chem* 1986;32:1341-6.
3. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol* 2006;59:798-801.
4. 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.
5. 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;54:729-37.
6. Siu AL, Bibbins-Domingo K, Grossman DC, et al. Screening for depression in adults: US Preventive Services Task Force recommendation statement. *JAMA* 2016;315:380-7.
7. Austin M, Highet N, the Expert Working Group. Mental Health Care in the Perinatal Period 2017.
8. Hewitt CE, Gilbody SM, Mann R, et al. Instruments to identify post-natal depression: Which methods have been the most extensively validated, in what setting and in which language?. *Int J Psychiatry Clin Pract* 2010;14:72-6.
9. Howard LM, Molyneaux E, Dennis CL, et al. Non-psychotic mental disorders in the perinatal period. *Lancet* 2014;384:1775-88.
10. 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. Submitted for review.
11. Pitkin RM, Branagan MA. Can the accuracy of abstracts be improved by providing specific instructions? A randomized controlled trial. *JAMA* 1998;280:267-9.

12. Beller EM, Glasziou PP, Altman DG, et al. PRISMA for Abstracts: reporting systematic reviews in journal and conference abstracts. *PLoS Med* 2013;10:e1001419.
13. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-III 3rd ed, revised. Washington, DC: American Psychiatric Association 1987.
14. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV 4th ed. Washington, DC: American Psychiatric Association 1994.
15. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV 4th ed, text revised. Washington, DC: American Psychiatric Association 2000.
16. World Health Organization. The ICD-10 Classifications of Mental and Behavioural Disorder: Clinical Descriptions and Diagnostic Guidelines. Geneva: World Health Organization 1992.
17. Thombs BD, Arthurs E, El-Baalbaki G, et al. Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review. *BMJ* 2011;343:d4825.
18. Sampson M, McGowan J, Lefebvre C, et al. PRESS: peer review of electronic search strategies. *Ottawa: Canadian Agency for Drugs and Technologies in Health* 2008.
19. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: development of the 10-item Edinburgh Postnatal Depression Scale. *Br J Psychiatry* 1987;150:782-6.
20. O'Connor E, Rossom RC, Henninger M, et al. Primary Care Screening for and Treatment of Depression in Pregnant and Postpartum Women: Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* 2016;315:388-406.
21. Hewitt C, Gilbody S, Brealey S, et al. Methods to identify postnatal depression in primary care: an integrated evidence synthesis and value of information analysis. *Health Technol Assess* 2009;13:1,145, 147.
22. 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.
23. 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.

24. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J* 2005;47:458-72.
25. Schisterman EF, Perkins N. Confidence Intervals for the Youden Index and Corresponding Optimal Cut-Point. *Commun Stat Simul Comput* 2007;36:549-63.
26. Smith GC, Seaman SR, Wood AM, et al. Correcting for optimistic prediction in small data sets. *Am J Epidemiol* 2014;180:318-24.
27. Thombs BD, Benedetti A, Kloda LA, et al. Diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for detecting major depression in pregnant and postnatal women: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open* 2015;5:e009742-2015.
28. 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.
29. Levis B, Benedetti A, Levis AW, et al. Selective cutoff reporting in studies of diagnostic test accuracy: a comparison of conventional and individual-patient-data meta-analyses of the Patient Health Questionnaire-9 depression screening tool. *Am J Epidemiol* 2017;185:954-64.
30. Dunn G, Pickles A, Tansella M, et al. Two-phase epidemiological surveys in psychiatric research. *Br J Psychiatry* 1999;174:95-100.
31. Thombs BD, Kwakkenbos L, Levis AW, et al. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. *CMAJ* 2018;190:E44-9.

Figure 1: Flow diagram of study selection process

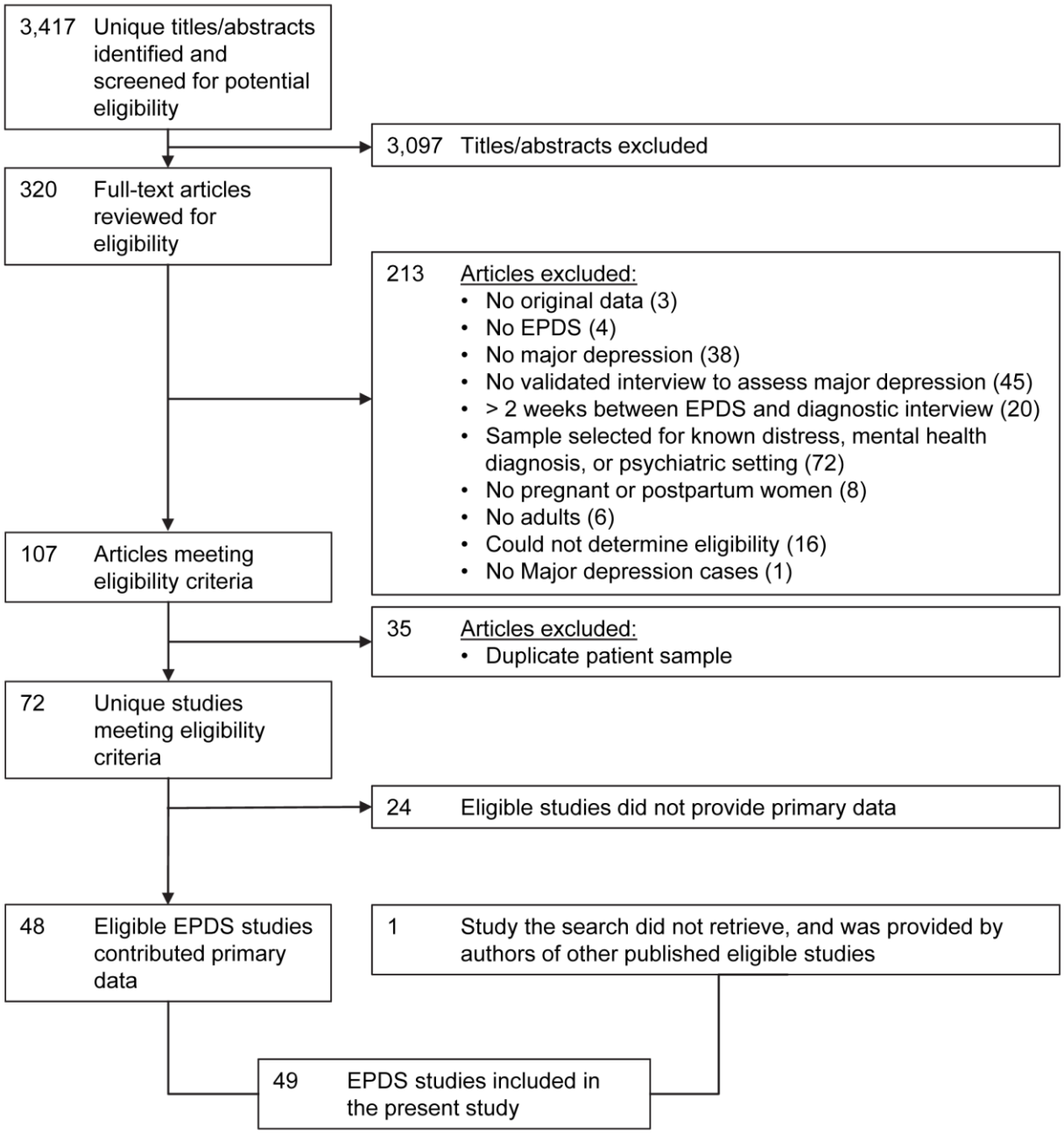
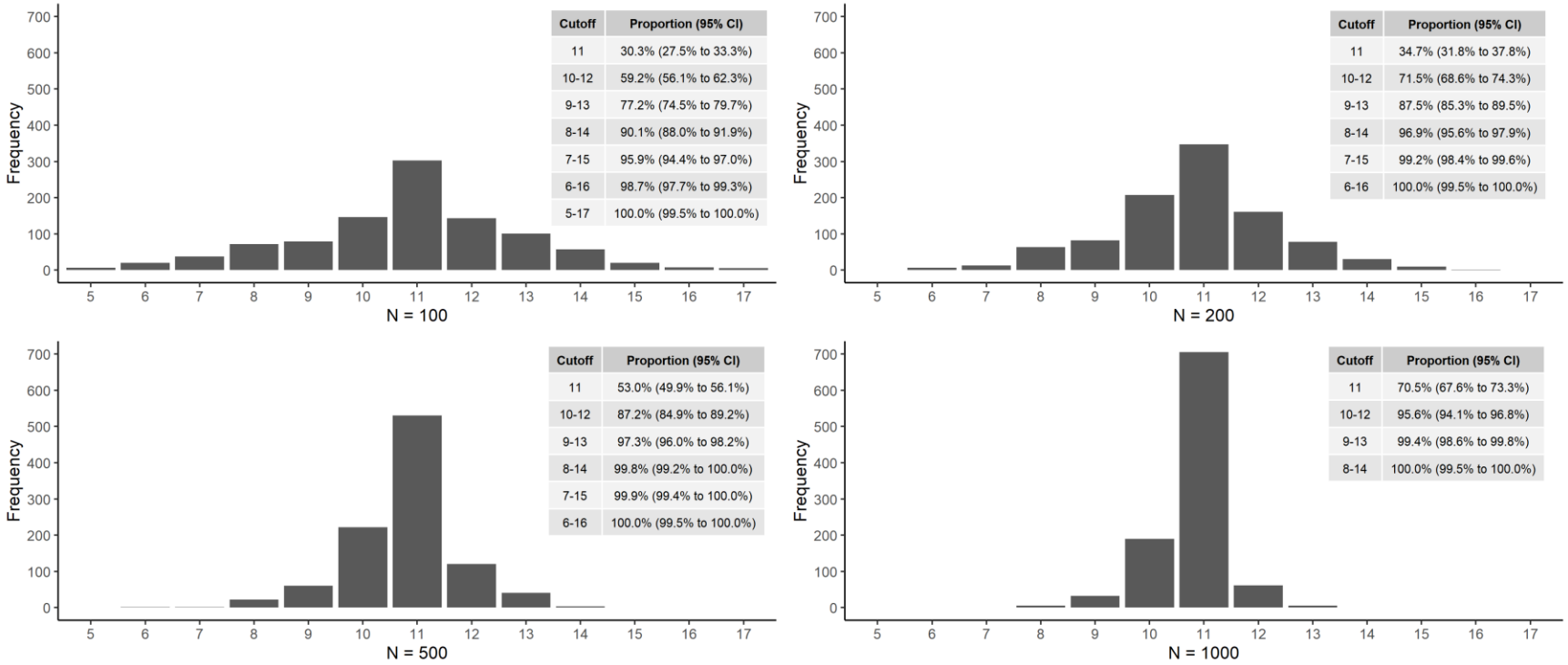
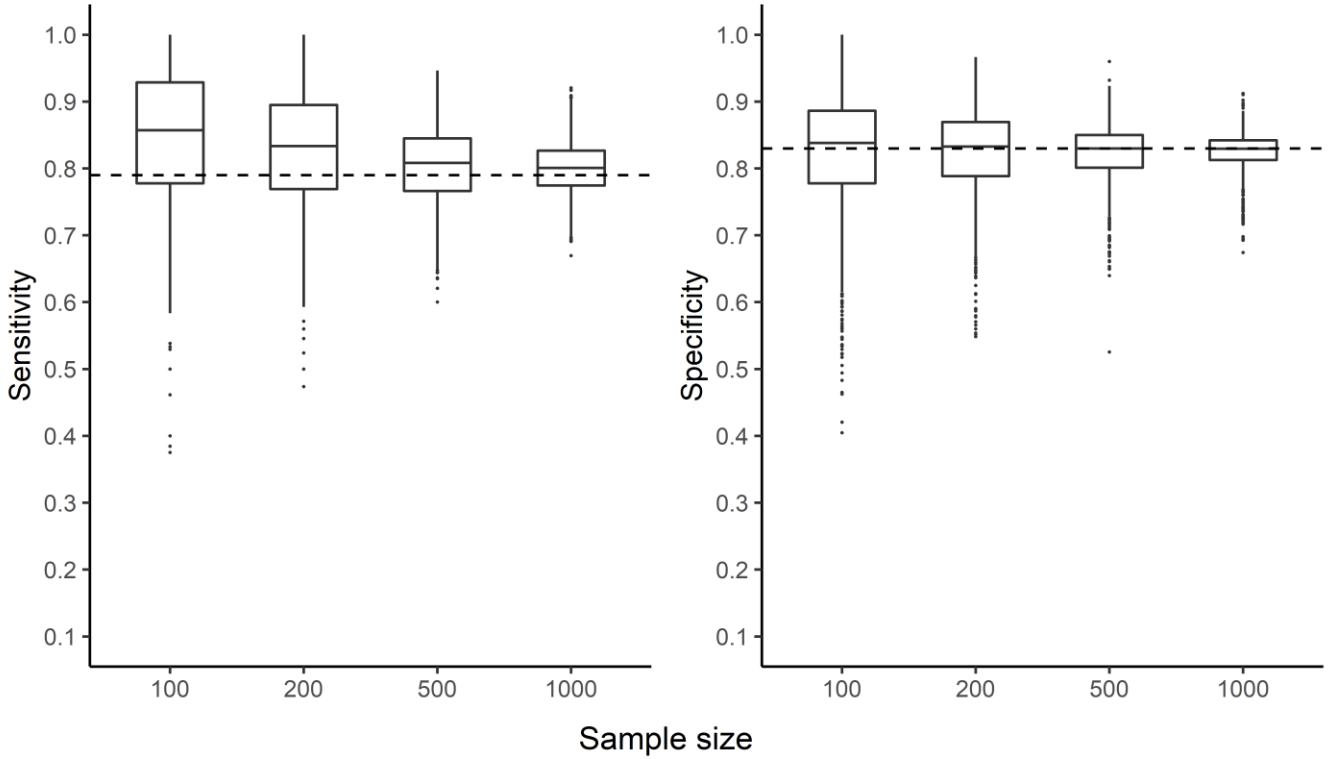


Figure 2: Variability in optimal cutoffs in 1,000 samples of size 100, 200, 500 and 1,000



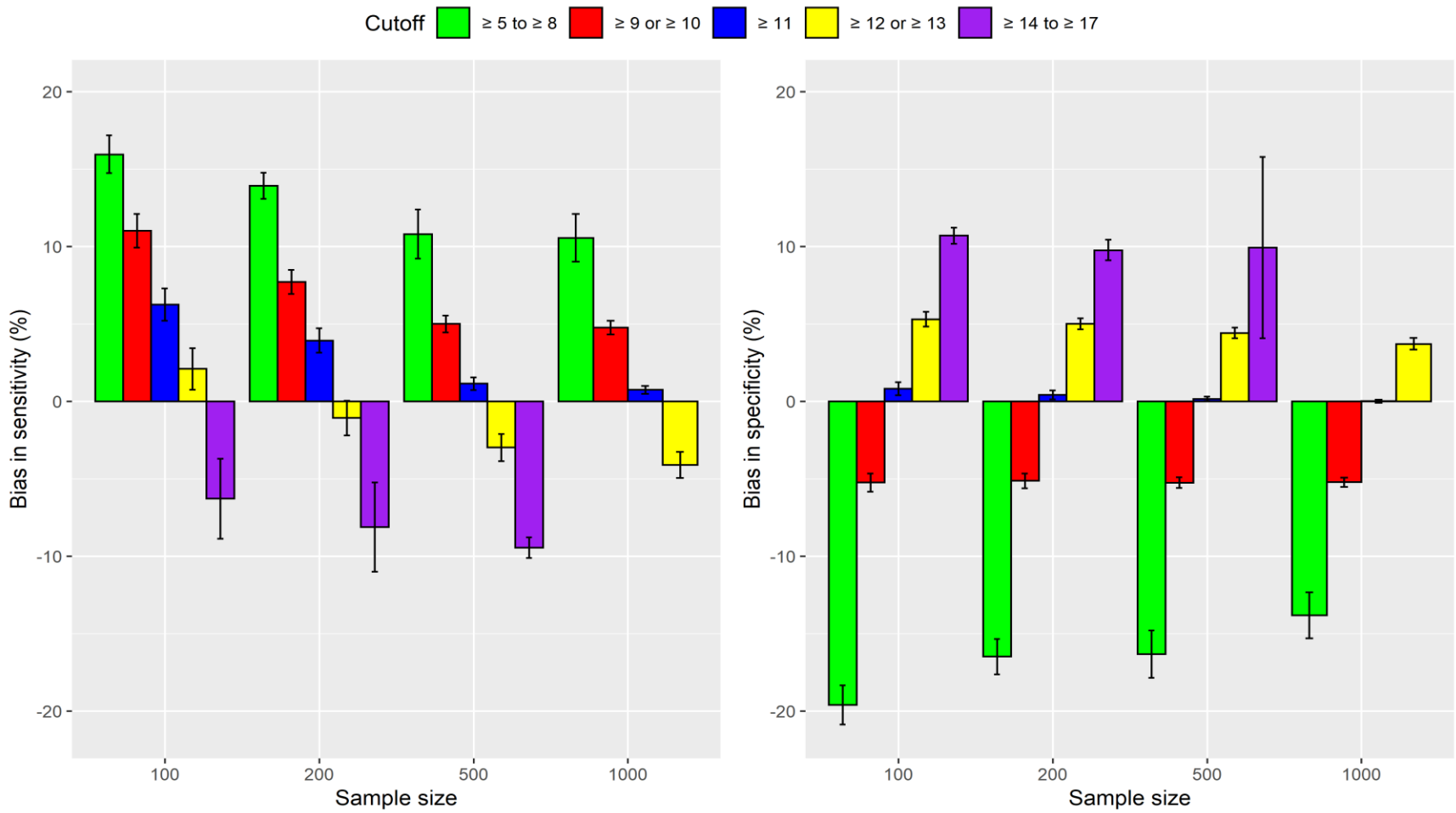
Optimal cutoff was defined as the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the study sample. 26 out of the 4,000 simulated samples had a tie in maximum Youden's J, and the higher cutoff was selected as the optimal cutoff.

Figure 3: Boxplots of accuracy estimates of the optimal cutoff in 1,000 samples of size 100, 200, 500 and 1,000 compared to the accuracy estimates of cutoff ≥ 11 in the population



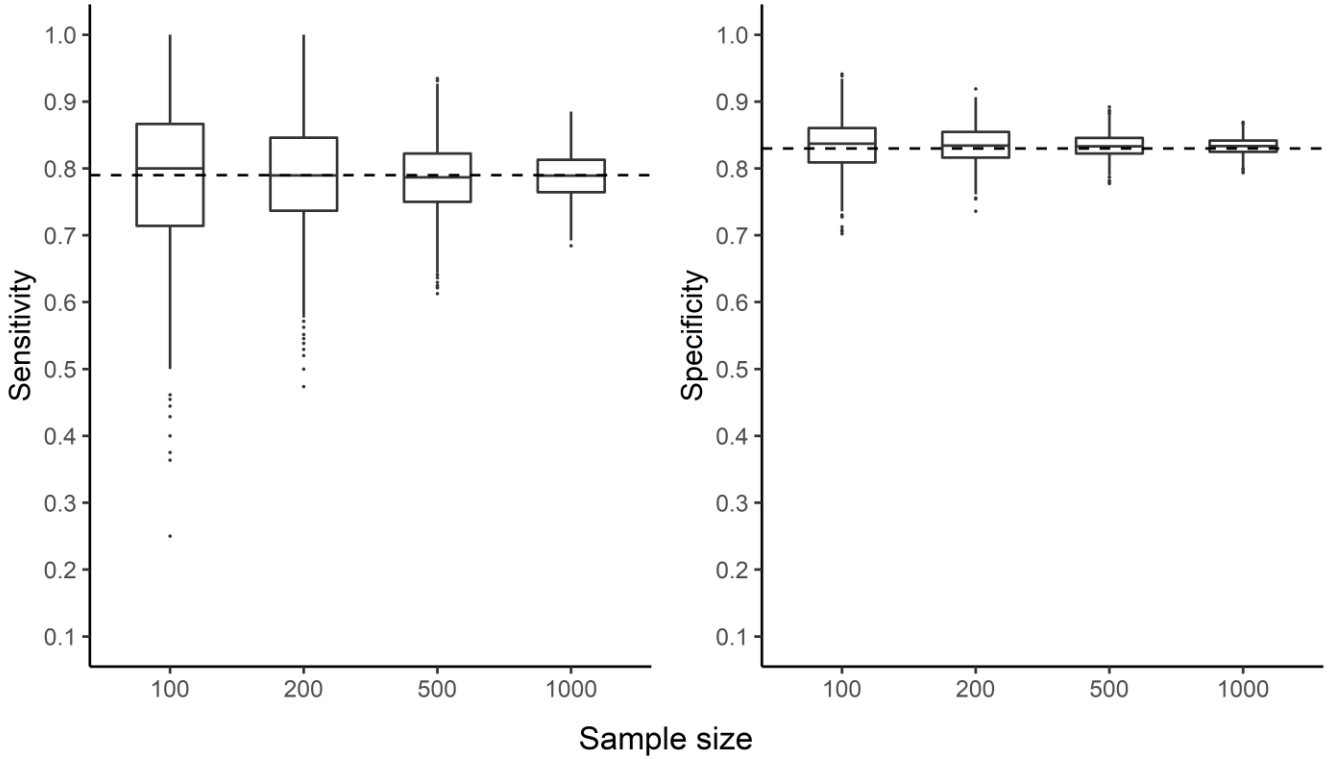
Optimal cutoff refers to the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the study sample. Dotted horizontal line represents the accuracy of cutoff ≥ 11 in the population (full IPDMA dataset).

Figure 4: Bias in accuracy estimates stratified across range of optimal cutoffs



Optimal cutoff refers to the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the study sample. The error bars represent 95% confidence intervals of the bias in accuracy estimates.

Figure 5: Boxplots of accuracy estimates of the cutoff ≥ 11 in 1,000 samples of size 100, 200, 500 and 1,000 compared to the accuracy estimates of cutoff ≥ 11 in the population



Dotted horizontal line represents the accuracy of cutoff ≥ 11 in the population (full IPDMA dataset).

Table 1: Bias in accuracy estimates in 1,000 samples of size 100, 200, 500 and 1,000

	Mean Difference (95% CI)							
	Sensitivity	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
	Sample size = 100		**Sample size = 200**		**Sample size = 500**		**Sample size = 1,000**	
Sample–specific optimal cutoff values – Population values with cutoff ≥ 11	6.5 (5.8 to 7.2)	-1.3 (-1.9 to -0.7)	4.2 (3.6 to 4.7)	-1.1 (-1.6 to -0.7)	1.8 (1.4 to 2.1)	-1.0 (-1.3 to -0.7)	1.4 (1.1 to 1.6)	-1.0 (-1.2 to -0.8)
Sample cutoff ≥ 11 values – Population values with cutoff ≥ 11	0.0 (-0.7 to 0.8)	0.1 (-0.2 to 0.3)	0.2 (-0.3 to 0.8)	0.1 (-0.1 to 0.3)	-0.2 (-0.6 to 0.1)	0.0 (-0.1 to 0.1)	0.1 (-0.1 to 0.3)	-0.1 (-0.1 to 0.0)

Optimal cutoff refers to the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the sample. Sample values are estimated from the simulated samples. Population values are estimated from the full dataset.

Chapter 4: Discussion

4.1 Thesis summary and implications

This thesis's objectives were to fill in the existing evidence gap, using actual participant data, on whether data-driven cutoff selection results in inaccurate optimal cutoff values and biases the accuracy estimates. For this purpose, a dataset originally synthesized for an IPDMA of diagnostic accuracy of EPDS to detect major depression was used.[15]

To address the thesis objective #1, variability in optimal cutoffs among study samples of different size was explored. It was found that data-driven cutoff selection resulted in selection of cutoffs that differed substantially from the population optimal cutoff established from the full IPDMA dataset. This result confirmed what was already reported in the literature about variability of data-driven optimal cutoffs,[5] with an actual participant data. In addition to confirming what was already known, this thesis is the first to document that the variability in optimal cutoffs is smaller in samples of larger sizes.

To address the thesis objective #2, accuracy estimates of optimal cutoffs estimated in study samples were compared with the true population accuracy estimates established from the full IPDMA dataset. It was found that sensitivity was substantially overestimated and specificity was minimally underestimated on average by cutoffs selected using data-driven technique, especially with smaller samples. This result confirmed the existing evidence on bias in accuracy estimates by cutoffs selected using data-driven techniques,[1-4] with an actual participant data.

A 2006 literature review of diagnostic accuracy studies published in eight "big" journals found that the median number of participants in a diagnostic accuracy study was 118 (interquartile range: 71 to 350 participants).[16] In the review of studies on diagnostic accuracy of EPDS too, studies often had small samples (Appendices: eMethods1). With the small sample size, the findings documented in this thesis suggest that most of these studies will likely estimate different optimal cutoff just because of random sample variation and will likely substantially overestimate sensitivity and minimally underestimate specificity.

Researchers conducting diagnostic accuracy studies should enroll larger samples and should recommend optimal cutoff only if their result is adequately cross-validated. Likewise, clinicians should select cutoff that is established from large samples and cross-validated across multiple studies. Ideally, the cutoff threshold for clinical use should be established from evidence synthesis of large numbers of well-conducted diagnostic accuracy studies.

4.2 Future research

There are several research questions that can follow-up on the research done for this thesis. First of all, it was the first time actual participant data was used to explore and quantify if cutoffs selected by data-driven approach are likely to be inaccurate and to produce biased accuracy estimates. It is crucial to explore if these results can be replicated with a different actual participant dataset. Second, in this thesis, data from EPDS test with ordinal scores was used and it will be interesting to see if these results can be replicated for other tests that measure symptoms in a continuous or ordinal scale. Third, Youden's J was selected as the method to optimize cutoff in this thesis. While it was found that Youden's J was frequently used to optimize and report cutoff in the abstract of EPDS diagnostic accuracy studies (Appendices: eMethods1), it is possible that a different method of optimizing cutoff would produce different results.

Finally, an important avenue of research would be to explore if the variability in cutoff selection and the bias in accuracy estimates from using cutoffs selected by data-driven approach can be statistically accounted for. Bootstrapping and several parametric or non-parametric methods to estimate confidence intervals for uncertainty around the optimal cutoff are available.[10,17] Similarly, internal validation methods such as cross-sampling, sample-splitting and bootstrapping could be used to statistically correct for the bias in accuracy estimates introduced by data-driven optimal cutoff selection.[18] Evaluating whether these methods can be successfully used to estimate the variability in optimal cutoff and correct for the bias in accuracy estimates from using data-driven cutoff selection approach would make an important contribution to the field of screening and diagnostic accuracy studies.

Chapter 5: Conclusions

In summary, this thesis provides evidence that optimal cutoffs identified by data-driven approach often diverge from the true population optimal cutoff and produce biased accuracy estimates, especially in samples of smaller size. Cutoff thresholds for use in clinical practice should be established from well-conducted meta-analyses and should only be recommended after cross-validation across multiple studies. Researchers should use larger samples to estimate optimal cutoff and diagnostic accuracy in their study sample. Future research with different measures, study samples and cutoff optimization methods should attempt to replicate the findings from this thesis. Also, research should focus on statistical methods to estimate the uncertainty in optimal cutoff and to correct the bias in accuracy estimates from using cutoffs estimated by data-driven approach.

Bibliography

1. Hirschfeld G, Zernikow B. Quantifying the variability of optimal cutpoints and reference values for diagnostic measures. *PeerJ PrePrints* 2014;2:e635v1.

2. Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin Chem* 1986;32(7):1341-6.

3. 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.

4. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol* 2006;59(8):798-801.

5. 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(4):449-53.

6. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019;38(11):2074-102.

7. InformedHealth.org [Internet]. Benefits and risks of screening tests Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2006 [updated 2019 Dec 17]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279418/ accessed Jan 16 2020.

8. Heleno B, Thomsen MF, Rodrigues DS, et al. Quantification of harms in cancer screening trials: literature review. *Bmj* 2013;347:f5334.

9. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3(1):32-5.

10. Schisterman EF, Perkins N. Confidence Intervals for the Youden Index and Corresponding Optimal Cut-Point. *Communications in Statistics – Simulation and Computation* 2007;36(3):549-63.

11. Boyd RC, Le HN, Somberg R. Review of screening instruments for postpartum depression. *Arch Womens Ment Health* 2005;8(3):141-53.

12. Howard LM, Molyneaux E, Dennis CL, et al. Non-psychotic mental disorders in the perinatal period. *Lancet* 2014;384(9956):1775-88.

13. National Collaborating Center for Mental Health. The NICE Guideline on the Management and Treatment of Depression in Adults (updated edition). London: NICE, 2010.

14. Siu AL, Bibbins-Domingo K, Grossman DC, et al. Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. *Jama* 2016;315(4):380-7.
15. Thombs BD, Benedetti A, Kloda LA, et al. Diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for detecting major depression in pregnant and postnatal women: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open* 2015;5(10):e009742.
16. Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. *Bmj* 2006;332(7550):1127-9.
17. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J* 2005;47(4):458-72.
18. Smith GC, Seaman SR, Wood AM, et al. Correcting for optimistic prediction in small data sets. *Am J Epidemiol* 2014;180(3):318-24.

Appendices

Data supplement for the manuscript in Chapter 3

eMethods1. Review of the cutoff reporting practice in the abstracts of studies on diagnostic accuracy of the EPDS

eMethods2. Search strategies

eFigure1. Distribution of optimal cutoffs for the 49 individual primary studies included in the IPDMA dataset

eTable1. Reasons for exclusion for all articles excluded at full-text level (N = 213)

eTable2. Characteristics of 49 primary studies included in the IPDMA dataset

eTable3. Characteristics of eligible primary studies that did not provide data for the present study (N = 24)

eTable4. Frequencies of EPDS scores for cases and non-cases of major depression in the full IPDMA dataset

eTable5. Bias in accuracy estimates of sample-specific optimal cutoffs compared to the accuracy estimates from population optimal cutoff, stratified by the magnitude of optimal cutoff

eMethods1. Review of the cutoff reporting practice in the abstracts of studies on diagnostic accuracy of the EPDS

OBJECTIVES

We carried out this review to explore cutoff reporting in the abstracts of studies on the diagnostic test accuracy of Edinburgh Postnatal Depression Scale (EPDS). We explored how often accuracy was reported for one versus multiple cutoffs. In studies where accuracy was reported for one cutoff only, we determined whether the cutoff was a data-driven optimal cutoff.

METHODS

Search strategy and identification of eligible studies

We searched PubMed for articles with records first listed in PubMed in 2014 or later using the search term ((depress*[Title] AND (sensitivity[Title/Abstract] OR specificity[Title/Abstract] OR accuracy[Title/Abstract]) AND (EPDS[Title/Abstract] OR edinburgh[Title/Abstract]))) AND ("2014"[Date - Create] : "3000"[Date - Create]). We chose an approximately 5-year period to capture recent practices.

Primary studies in any language that reported in the abstract the sensitivity and specificity of at least one cutoff on the Edinburgh Postnatal Depression Scale (EPDS) for detecting depression based on a diagnosis of major depression or a combination of depressive disorders were included. Search results were uploaded to DistillerSR (Evidence Partners, Ottawa, Canada), and two investigators independently reviewed the titles and abstracts for eligibility. If either investigator determined that a study was potentially eligible based on title and abstract, it underwent full-text review by two investigators independently. Conflicts between reviewers after full-text review were resolved by consensus, including a third investigator as necessary.

Evaluation of eligible studies

One investigator extracted data from the eligible studies into DistillerSR, and data were validated by a second investigator using the DistillerSR Quality Control function. Conflicts were resolved by consensus, involving a third investigator, as necessary. In addition to the information on the cutoffs for which accuracy was reported in the abstract, we also collected information on study population, number of study participants assessed

with both the EPDS and the diagnostic interview and the number of cases of depression for each study.

RESULTS

We identified 78 unique citations from the PubMed search. Out of the 78 citations, 64 did not meet eligibility criteria and were excluded. We extracted information from the abstracts of the remaining 14 studies. Of the 14 studies, only one (7%) reported accuracy estimates for multiple cutoffs in the abstract, and 13 (93%) reported accuracy for the study defined optimal cutoff only (table below).

Studies included in the review of cutoff reporting practice in studies on EPDS diagnostic accuracy

Pubmed ID	Author	Journal	Year	Sample size	Cutoff(s) reported	Type of cutoff(s)
24733682	Baillon[1]	Int J Geriatr Psychiatry	2014	120	≥ 11	Optimal
24742635	Toreki[2]	Midwifery	2014	266	≥ 8 and ≥ 13	Cutoff ≥ 8 and optimal
25110542	Alvarado-Esquivel[3]	J Clin Med Res	2014	158	≥ 10	Optimal
25293375	Matijasevich[4]	BMC Psychiatry	2014	447	≥ 8	Optimal
25380783	Figueiredo[5]	Arch Womens Ment Health	2015	199	≥ 10	Optimal
25493092	Alvarado-Esquivel[6]	Clin Pract Epidemiol Ment Health	2014	120	≥ 9	Optimal
25754606	Martins[7]	J Affect Disord	2015	807	≥ 10	Optimal
25770478	Castro E Couto[8]	J Affect Disord	2015	Not reported	≥ 11	Optimal
26185471	Khalifa[9]	Int J Womens Health	2015	Not reported	≥ 12	Optimal
27126326	Esiwe[10]	Age Ageing	2016	118	≥ 8	Optimal
27785152	Bhusal[11]	Int J Ment Health Syst	2016	346	≥ 13	Optimal
28767198	Usuda[12]	Psychiatry Clin Neurosci	2017	210	≥ 13	Optimal
30396159	Lydsdottir[13]	Midwifery	2019	474	≥ 11	Optimal

30599376	Vazquez[14]	J Affect Disord	2019	569	≥ 10	Optimal

Optimal cutoff refers to cutoff that was reported as optimal, best-performing, most accurate, maximizing Youden's J or in similar terms by the authors in the study.

eMethods1 References

1. Baillon S, Lindesay J, Prabhakaran P, Hands O, Murray J, Stacey S, et al. The utility of the Edinburgh Depression Scale as a screening tool for depression in Parkinson's disease. *Int J Geriatr Psychiatry* 2014;29:1286-93.

2. Toreki A, Andó B, Dudas RB, Dweik D, Janka Z, Kozinszky Z, et al. Validation of the Edinburgh Postnatal Depression Scale as a screening tool for postpartum depression in a clinical sample in Hungary. *Midwifery* 2014;30:911-8.

3. Alvarado-Esquivel C, Sifuentes-Alvarez A, Salas-Martinez C. Validation of the edinburgh postpartum depression scale in a population of adult pregnant women in Mexico. *J Clin Med Res* 2014;6:374-8.

4. Matijasevich A, Munhoz TN, Tavares BF, Barbosa AP, da Silva DM, Abitante MS, et al. Validation of the Edinburgh Postnatal Depression Scale (EPDS) for screening of major depressive episode among adults from the general population. *BMC Psychiatry* 2014;14:284.

5. Figueiredo FP, Parada AP, Cardoso VC, Batista RF, Silva AA, Barbieri MA, et al. Postpartum depression screening by telephone: a good alternative for public health and research. *Arch Womens Ment Health* 2015;18:547-53.

6. Alvarado-Esquivel C, Sifuentes-Alvarez A, Salas-Martinez C. The use of the edinburgh postpartum depression scale in a population of teenager pregnant women in Mexico: a validation study. *Clin Pract Epidemiol Ment Health* 2014;10:129-32.

7. Martins Cde S, Motta JV, Quevedo LA, Matos MB, Pinheiro KA, Souza LD, et al. Comparison of two instruments to track depression symptoms during pregnancy in a sample of pregnant teenagers in Southern Brazil. *J Affect Disord* 2015;177:95-100.

8. Castro E Couto T, Martins Brancaglion MY, Nogueira Cardoso M, Bergo Protzner A, Duarte Garcia F, Nicolato R, et al. What is the best tool for screening antenatal depression. *J Affect Disord* 2015;178:12-7.

9. Khalifa DS, Glavin K, Bjertness E, Lien L. Postnatal depression among Sudanese women: prevalence and validation of the Edinburgh Postnatal Depression Scale at 3 months postpartum. *Int J Womens Health* 2015;7:677-84.

10. Esiwe C, Baillon S, Rajkonwar A, Lindesay J, Lo N, Dennis M. Screening for depression in older people on acute medical wards: the validity of the Edinburgh Depression Scale. *Age Ageing* 2016;45:554-8.

11. Bhusal BR, Bhandari N, Chapagai M, Gavidia T. Validating the Edinburgh Postnatal Depression Scale as a screening tool for postpartum depression in Kathmandu, Nepal. *Int J Ment Health Syst* 2016;10:71.

12. Usuda K, Nishi D, Okazaki E, Makino M, Sano Y. Optimal cut-off score of the Edinburgh Postnatal Depression Scale for major depressive episode during pregnancy in Japan. *Psychiatry Clin Neurosci* 2017;71:836-42.

13. Lydsdottir LB, Howard LM, Olafsdottir H, Thome M, Tyrfingsson P, Sigurdsson JF. The psychometric properties of the Icelandic version of the Edinburgh Postnatal Depression Scale (EPDS) when used prenatal. *Midwifery* 2019;69:45-51.

14. Vázquez MB, Míguez MC. Validation of the Edinburgh postnatal depression scale as a screening tool for depression in Spanish pregnant women. *J Affect Disord* 2019;246:515-21.

eMethods2. Search strategies

MEDLINE (OvidSP)

1. EPDS.af.
2. Edinburgh Postnatal Depression.af.
3. Edinburgh Depression Scale.af.
4. or/1-3
5. Mass Screening/
6. Psychiatric Status Rating Scales/
7. "Predictive Value of Tests"/
8. "Reproducibility of Results"/
9. exp "Sensitivity and Specificity"/
10. Psychometrics/
11. Prevalence/
12. Reference Values/
13. Reference Standards/
14. exp Diagnostic Errors/
15. Mental Disorders/di, pc [Diagnosis, Prevention & Control]
16. Mood Disorders/di, pc [Diagnosis, Prevention & Control]
17. Depressive Disorder/di, pc [Diagnosis, Prevention & Control]
18. Depressive Disorder, Major/di, pc [Diagnosis, Prevention & Control]
19. Depression, Postpartum/di, pc [Diagnosis, Prevention & Control]
20. Depression/di, pc [Diagnosis, Prevention & Control]
21. validation studies.pt.
22. comparative study.pt.
23. screen*.af.
24. prevalence.af.
25. predictive value*.af.
26. detect*.ti.
27. sensitiv*.ti.
28. valid*.ti.

29. revalid*.ti.

30. predict*.ti.

31. accura*.ti.

32. psychometric*.ti.

33. identif*.ti.

34. specificit*.ab.

35. cut?off*.ab.

36. cut* score*.ab.

37. cut?point*.ab.

38. threshold score*.ab.

39. reference standard*.ab.

40. reference test*.ab.

41. index test*.ab.

42. gold standard.ab.

43. or/5-42

44. 4 and 43

PsycINFO (OvidSP)

1. EPDS.af.
2. Edinburgh Postnatal Depression.af.
3. Edinburgh Depression Scale.af.
4. or/1-3
5. Diagnosis/
6. Medical Diagnosis/
7. Psychodiagnosis/
8. Misdiagnosis/
9. Screening/
10. Health Screening/
11. Screening Tests/
12. Prediction/
13. Cutting Scores/
14. Psychometrics/
15. Test Validity/
16. screen*.af.
17. predictive value*.af.
18. detect*.ti.
19. sensitiv*.ti.
20. valid*.ti.
21. revalid*.ti.
22. accura*.ti.
23. psychometric*.ti.
24. specificit*.ab.
25. cut?off*.ab.
26. cut* score*.ab.
27. cut?point*.ab.
28. threshold score*.ab.
29. reference standard*.ab.

30. reference test*.ab.

31. index test*.ab.

32. gold standard.ab.

33. or/5-32

34. 4 and 33

Web of Science (Web of Knowledge)

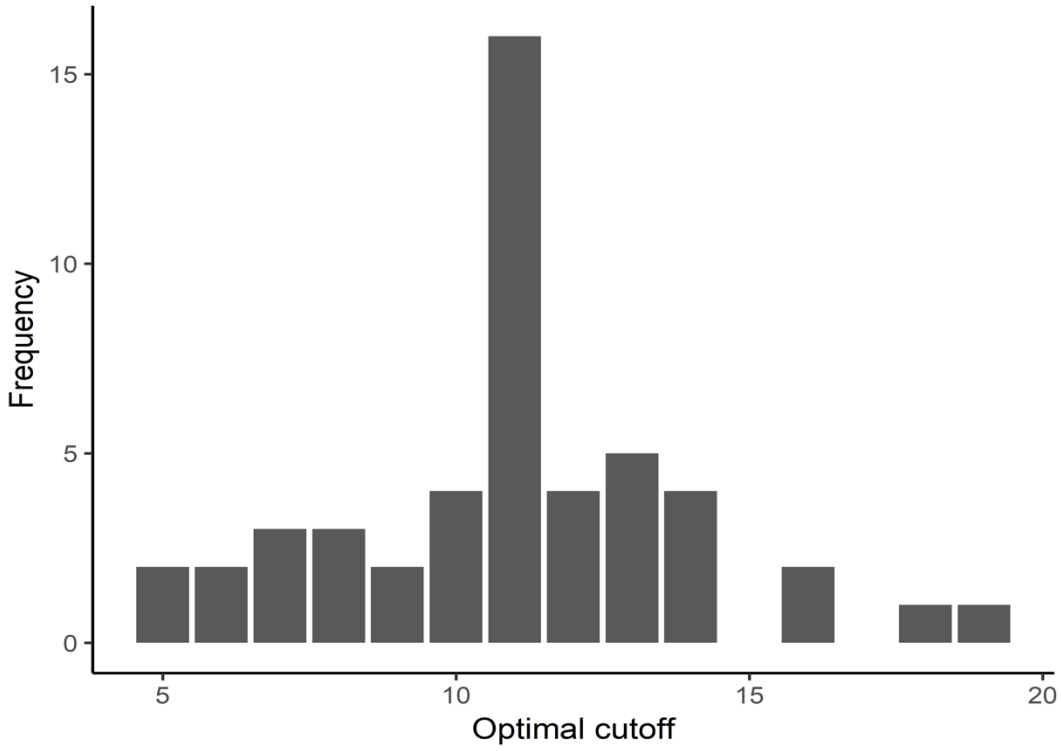
#1. TS=(EPDS OR “Edinburgh Postnatal Depression” OR “Edinburgh Depression Scale”)

#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” OR “reliab*”)

#2 AND #1

Databases=SCI-EXPANDED, SSCI, A&HCI

eFigure1. Distribution of optimal cutoffs for the 49 individual primary studies included in the IPDMA dataset



Optimal cutoff refers to the cutoff that maximized Youden's J (sensitivity + specificity - 1) in each primary study dataset.

eTable1. Reasons for exclusion for all articles excluded at full–text level (N = 213)

Reference	Reason for exclusion
Abiodun OA. Postnatal depression in primary care populations in Nigeria. Gen Hosp Psychiatry. 2006;**28**:133-6.	Could not determine eligibility[a]
Abou-Saleh MT, Ghubash R, Karim L, Krymski M, Bhai I. Hormonal aspects of postpartum depression. Psychoneuroendocrinology. 1998;**23**:465-75.	> 2 weeks between EPDS and diagnostic interview
Aceti F, Baglioni V, Ciolli P, De Bei F, Di Lorenzo F, Ferracuti S, et al. Maternal attachment patterns and personality in post partum depression. Riv Psichiatr. 2012;**47**:214-20.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Adewuya AO, Eegunranti AB, Lawal AM. Prevalence of postnatal depression in Western Nigerian women: a controlled study. Int J Psychiatry Clin Pract. 2005;**9**:60-4.	Could not determine eligibility[a]
Adewuya AO. Early postpartum mood as a risk factor for postnatal depression in Nigerian women. Am J Psychiatry. 2006;**163**:1435-7.	No validated interview to assess major depression
Ahn S, Corwin EJ. The association between breastfeeding, the stress response, inflammation, and postpartum depression during the postpartum period: Prospective cohort study. Int J Nurs Stud. 2015;**52**:1582-90.	Major depression not assessed
Alami KM, Kadri N, Berrada S. Prevalence and psychosocial correlates of depressed mood during pregnancy and after childbirth in a Moroccan sample. Arch Womens Ment HealthArch Womens Ment Health. 2006;**9**:343-6.	Could not determine eligibility[a]
Albacar G, Sans T, MartinSantos R, GarciaEsteve L, Guillamat R, Sanjuan J, et al. Thyroid function 48 h after delivery as a marker for subsequent postpartum depression. Psychoneuroendocrinology. 2010;**35**:738-42.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Albacar G, Sans T, MartinSantos R, GarciaEsteve L, Guillamat R, Sanjuan J, et al. An association between plasma ferritin concentrations measured 48h after delivery and postpartum depression. J Affect DisordJ Affect Disord. 2011;**131**:136-42.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Alexander S, Palmer C, Stone PC. Evaluation of screening instruments for depression and anxiety in breast cancer survivors. Breast Cancer Res Treat. 2010;**122**:573-8.	No pregnant or postpartum women
Algul A, Semiz UB, Dundar O, Ates MA, Basoglu C, Ebrinc S, et al. Psychosocial and hormone related risk factors for early postnatal depressive symptoms in Turkish women. Neurol Psychiat Br. 2008;**15**:117-22.	Major depression not assessed
Al-Modayfer O, Alatiq Y, Khair O, Abdelkawi S. Postpartum depression and related risk factors among Saudi females. Int J Cult Ment Health. 2015;**8**:316-24.	No validated interview to assess major depression
Alvarado-Esquivel C, Sifuentes-Alvarez A, Estrada-Martínez S, Salas-Martínez C, Hernndez-Alvarado AB, Ortiz-Rocha SG, et al. Prevalence of postnatal depression in women attending public hospitals in Durango, Mexico. Gac Med Mex. 2010;**146**:1-9.	No validated interview to assess major depression
Alvarado-Esquivel C, Sifuentes-Alvarez A, Salas-Martinez C. Unhappiness with the Fetal Gender is associated with Depression in Adult Pregnant Women Attending Prenatal Care in a Public Hospital in Durango, Mexico. Int J Biomed Sci. 2016;**12**:36-41.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Areias ME, Kumar R, Barros H, Figueiredo E. Comparative incidence of depression in women and men, during pregnancy and after childbirth. Validation of the Edinburgh Postnatal Depression Scale in Portuguese mothers. Br J Psychiatry. 1996;**169**:30-5.	No validated interview to assess major depression
Areias ME, Kumar R, Barros H, Figueiredo E. Correlates of postnatal depression in mothers and fathers. Br J Psychiatry. 1996;**169**:36-41.	No validated interview to assess major depression

Austin MP, Dudley M, Launders C, Dixon C, Macartney-Bourne F. Description and evaluation of a domiciliary perinatal mental health service focussing on early intervention. Arch Womens Ment Health. 1999;**2**:169-73.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Austin MP, Frilingos M, Lumley J, Hadzi-Pavlovic D, Roncolato W, Acland S, et al. Brief antenatal cognitive behaviour therapy group intervention for the prevention of postnatal depression and anxiety: a randomised controlled trial. J Affect Disord. 2008;**105**:35-44.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Austin MP, Hadzi-Pavlovic D, Priest SR, Reilly N, Wilhelm K, Saint K, Parker G. Depressive and anxiety disorders in the postpartum period: how prevalent are they and can we improve their detection? Arch Womens Ment Health. 2010;**13**:395-401.	Major depression not assessed
Austin MP, Hadzi-Pavlovic D, Saint K, Parker G. Antenatal screening for the prediction of postnatal depression: validation of a psychosocial Pregnancy Risk Questionnaire. Acta Psychiatr Scand. 2005;**112**:310-7.	Major depression not assessed
Azar R, Paquette D, Zoccolillo M, Baltzer F, Tremblay RE. The association of major depression, conduct disorder, and maternal overcontrol with a failure to show a cortisol buffered response in 4-month-old infants of teenage mothers. Biological Psychiatry. 2007;**62**:573-9.	Not a sample of adults
Bågedahl□Strindlund M, Monsen Börjesson K. Postnatal depression: a hidden illness. Acta Psychiatr Scand. 1998;**98**:272-5.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Bergant AM, Heim K, Ulmer H, Illmensee K. Early postnatal depressive mood: associations with obstetric and psychosocial factors. J Psychosom Res. 1999;**46**:391-4.	Major depression not assessed

Bergant AM, Nguyen T, Heim K, Ulmer H, Dapunt O. German language version and validation of the Edinburgh postnatal depression scale. Dtsch Med Wochenschr. 1998;**123**:35-40.	No validated interview to assess major depression
Bick DE, MacArthur C, Lancashire RJ. What influences the uptake and early cessation of breast feeding? Midwifery. 1998;**14**:242-7.	Major depression not assessed
Bloch M, Rotenberg N, Koren D, Klein E. Risk factors associated with the development of postpartum mood disorders. J Affect Disord. 2005;**88**:9-18.	> 2 weeks between EPDS and diagnostic interview
Boath E, Cox J, Lewis M, Jones P, Pryce A. When the cradle falls: the treatment of postnatal depression in a psychiatric day hospital compared with routine primary care. J Affect Disord. 1999;**53**:143-51.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Boyce P, Hickey A. Psychosocial risk factors to major depression after childbirth. Soc Psychiatry Psychiatr Epidemiol. 2005;**40**:605-12.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Boyce P, Stubbs J, Todd A. The Edinburgh Postnatal Depression Scale: validation for an Australian sample. Aust N Z J Psychiatry. 1993;**27**:472-6.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Browne JC, Scott KM, Silvers KM. Fish consumption in pregnancy and omega-3 status after birth are not associated with postnatal depression. J Affect Disord. 2006;**90**:131-9.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Brugha TS, Wheatley S, Taub NA, Culverwell A, Friedman T, Kirwan P, et al. Pragmatic randomized trial of antenatal intervention to prevent post-natal depression by reducing psychosocial risk factors. Psychol Med. 2000;**30**:1273-81.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Buneviius A, Kusminskas L, Buneviius R. Validation of the Lithuanian version of the Edinburgh Postnatal Depression Scale. Med Lith. 2009;**45**:544.	No validated interview to assess major depression
Bunevicius A, Kusminskas L, Bunevicius R. Validity of the Edinburgh Postnatal Depression Scale. Eur Psychiatry. 2009;**24:**S896.	No validated interview to assess major depression

Burns A, O'Mahen H, Baxter H, Bennert K, Wiles N, Ramchandani P, et al. A pilot randomised controlled trial of cognitive behavioural therapy for antenatal depression. BMC Psychiatry. 2013;**13**:33.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Byatt N, Biebel K, Simas TAM, Sarvet B, Ravech M, Allison J, Straus J. Improving perinatal depression care: The Massachusetts Child Psychiatry Access Project for Moms. Gen Hosp Psychiatry. 2016;**40**:12-7.	Major depression not assessed
Caramlau I, Barlow J, Sembi S, McKenzie-McHarg K, McCabe C. Mums 4 Mums: structured telephone peer-support for women experiencing postnatal depression. Pilot and exploratory RCT of its clinical and cost effectiveness. Trials . 2011;**12**:88.	No original data
Carothers AD, Murray L. Estimating psychiatric morbidity by logistic regression: application to post-natal depression in a community sample. Psychol Med. 1990;**20**:695-702.	No validated interview to assess major depression
Carpiniello B, Pariante CM, Serri F, Costa G, Carta MG. Validation of the Edinburgh Postnatal Depression Scale in Italy. J Psychosom Obstet Gynecol. 1997;**18**:280-5.	No validated interview to assess major depression
Castañón SC, Pinto LJ. Use of the Edinburgh Postnatal Depression Scale to detect postpartum depression. Rev Med Chil. 2008;**136**:851-8.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Chaudron LH, Nirodi N. The obsessive-compulsive spectrum in the perinatal period: a prospective pilot study. Arch Womens Ment Health. 2010;**13**:403-10.	> 2 weeks between EPDS and diagnostic interview
Chee CY, Chong YS, Ng TP, Lee DT, Tan LK, Fones CS. The association between maternal depression and frequent non-routine visits to the infant's doctor--a cohort study. J Affect Disord. 2008;**107**:247-53.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Chee CYI, Lee DTS, Chong YS, Tan LK, Ng TR, Fones CSL. Confinement and other psychosocial factors in perinatal depression: A transcultural study in Singapore. J Affect Disord. 2005;**89**:157-66.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Chen H, Bautista D, Ch'ng YC, Li W, Chan E, Rush AJ. Screening for postnatal depression in Chinese-speaking women using the Hong Kong translated version of the Edinburgh Postnatal Depression Scale. Asia Pac Psychiatry. 2013;**5**:E64-E72.	No validated interview to assess major depression
Chibanda D, Verhey R, Gibson LJ, Munetsi E, Machando D, Rusakaniko S, 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.	EPDS not administered
Clarke PJ. Validation of two postpartum depression screening scales with a sample of First Nations and Metis women. Can J Nurs Res. 2008;**40**:112-25.	Major depression not assessed
Class QA, Verhulst J, Heiman JR. Exploring the heterogeneity in clinical presentation and functional impairment of postpartum depression. J Reprod Infant Psychol. 2013;**31**:183-94	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Clifford C, Day A, Cox J, Werrett J. A cross-cultural analysis of the use of the Edinburgh Post-Natal Depression Scale (EPDS) in health visiting practice. J Adv Nurs. 1999;**30**:655-64.	No validated interview to assess major depression
Coleman R, Morison L, Paine K, Powell RA, Walraven G. Women's reproductive health and depression: a community survey in the Gambia, West Africa. Soc Psychiatry Psychiatr Epidemiol. 2006;**41**:720-7.	No validated interview to assess major depression
Cooper PJ, Murray L, Wilson A, Romaniuk H. Controlled trial of the short- and long-term effect of psychological treatment of post-partum depression. I. Impact on maternal mood. Br J Psychiatry. 2003;**182**:412-9.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Costas J, Gratacòs M, Escaramís G, Martín-Santos R, de Diego Y, Baca- García E, et al. Association study of 44 candidate genes with depressive and anxiety symptoms in post-partum women. J Psychiatr Res. 2010;**44**:717-24.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Cox JL, Chapman G, Murray D, Jones P. Validation of the Edinburgh Postnatal Depression Scale (EPDS) in non-postnatal women. J Affect Disord. 1996;**39**:185-9.	No validated interview to assess major depression
Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry. 1987;**150**:782-6.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Cox JL, Murray D, Chapman G. A controlled study of the onset, duration and prevalence of postnatal depression. Br J Psychiatry. 1993;**163**:27-31.	No validated interview to assess major depression
de Souza Ribeiro Martins C, dos Santos Motta JV, Quevedo LA, de Matos MB, Pinheiro KAT, de Mattos Souza LD, et al. Comparison of two instruments to track depression symptoms during pregnancy in a sample of pregnant teenagers in Southern Brazil. J Affect Disord. 2015;**177**:95-100.	Not a sample of adults
Dennis CL, Hodnett E, Kenton L, Weston J, Zupancic J, Stewart DE, Kiss A. Effect of peer support on prevention of postnatal depression among high risk women: multisite randomised controlled trial. BMJ. 2009;**338**:a3064.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Ebeigbe PN, Akhigbe KO. Incidence and associated risk factors of postpartum depression in a tertiary hospital in Nigeria. Niger Postgrad Med J. 2008;**15**:15-8.	Major depression not assessed
Eberhard-Gran M, Eskild A, Tambs K, Schei B, Opjordsmoen S. The Edinburgh Postnatal Depression Scale: validation in a Norwegian community sample. Nord J Psychiatry. 2001;**55**:113-7.	No validated interview to assess major depression
Ekeroma AJ, Ikenasio-Thorpe B, Weeks S, Kokaua J, Puniani K, Stone P, Foliaki SA. Validation of the Edinburgh Postnatal Depression Scale (EPDS) as a screening tool for	> 2 weeks between EPDS and diagnostic interview

postnatal depression in Samoan and Tongan women living in New Zealand. N Z M J. 2012;**125**:41-9.	
Ekuklu G, Tokuc B, Eskiocak M, Berberoglu U, Saltik A. Prevalence of postpartum depression in Edirne, Turkey, and related factors. J Reprod Med. 2004;**49**:908-14.	Major depression not assessed
El-Ibiary SY, Hamilton SP, Abel R, Erdman CA, Robertson PA, Finley PR. A pilot study evaluating genetic and environmental factors for postpartum depression. Innov Clin Neurosci. 2013;**10**:15-22.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Elliott SA, Leverton TJ, Sanjack M, Turner H, Cowmeadow P, Hopkins J, Bushnell D. Promoting mental health after childbirth: a controlled trial of primary prevention of postnatal depression. Br J Clin Psychol. 2000;**39**:223-41.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Fairbrother N, Young AH, Janssen P, Antony MM, Tucker E. Depression and anxiety during the perinatal period. BMC Psychiatry. 2015;**15**:206.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Farhat A, Saeidi R, Mohammadzadeh A, Hesari H. Prevalence of Postpartum Depression; a longitudinal study. Iran J Neonatol. 2015;**6**:39-44.	Major depression not assessed
Flynn HA, Sexton M, Ratliff S, Porter K, Zivin K. Comparative performance of the Edinburgh Postnatal Depression Scale and the Patient Health Questionnaire-9 in pregnant and postpartum women seeking Psychiatr Serv. Psychiatry Res. 2011;**187**:130-4.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gallanti AME, Rodríguez CEAM, Rodríguez IM, Sosa MA. Puerperal depression and its association with demographic and social factors, the way of resolution of pregnancy and the newborn clinical evolution. Medula. 2015;**24**:25-34.	Major depression not assessed

Gelabert E, Subira S, Plaza A, Torres A, Navarro P, Imaz ML, et al. The Vulnerable Personality Style Questionnaire: psychometric properties in Spanish postpartum women. Arch Womens Ment Health. 2011;**14**:115-24.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gemmill AW, Leigh B, Ericksen J, Milgrom J. A survey of the clinical acceptability of screening for postnatal depression in depressed and non-depressed women. BMC Public Health. 2006;**6**:211.	Major depression not assessed
Georgiopoulos AM, Bryan TL, Wollan P, Yawn BP. Routine screening for postpartum depression. J Fam Pract. 2001;**50**:117.	Major depression not assessed
Gerardin P, Wendland J, Bodeau N, Galin A, Bialobos S, Tordjman S, et al. Depression during pregnancy: Is the developmental impact earlier in boys? A prospective case-control study. J Clin Psychiatry. 2011;**72**:378-87.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gerardin P. Characteristics and clinical consequences of prenatal depression. Main results of a prospective case-control study on perinatal depression from pregnancy to one year-old infant. Neuropsychiatr Enfance eAdolesc. 2012;**60**:138-46.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Ghubash R, Abou-Saleh MT, Daradkeh TK. The validity of the Arabic Edinburgh Postnatal Depression Scale. Soc Psychiatry Psychiatr Epidemiol. 1997;**32**:474-6.	> 2 weeks between EPDS and diagnostic interview
Ghubash R, Abou-Saleh MT. Postpartum psychiatric illness in Arab culture: prevalence and psychosocial correlates. Br J Psychiatry. 1997;**171**:65-8.	> 2 weeks between EPDS and diagnostic interview
Ginsburg GS, Barlow A, Goklish N, Hastings R, Baker EV, Mullany B, et al. Postpartum depression prevention for reservation-based American Indians: Results from a Pilot Randomized Controlled Trial. Child Youth Care Forum. 2012;**41**:229-45.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Goeb JL, Férel S, Guetta J, Guibert J, Guedeney A, Coste J, et al. Assisted reproductive techniques when the Man is HIV Seropositive. Psychiatr Enfant. 2009;**52**:63-88.	Major depression not assessed

Goutaudier N, Lopez A, SéjournéN, Denis A, Chabrol H. Premature birth: subjective and psychological experiences in the first weeks following childbirth, a mixed-methods study. J Reprod Infant Psychol. 2011;**29**:364-73.	Major depression not assessed
Goyal D, Park VT, McNiesh S. Postpartum depression among Asian Indian mothers. MCN Am J Matern Child Nurs. 2015;**40**:256-61.	Major depression not assessed
Grant KA, Bautovich A, McMahon C, Reilly N, Leader L, Austin MP. Parental care and control during childhood: Associations with maternal perinatal mood disturbance and parenting stress. Arch Womens Ment Health. 2012;**15**:297-305.	Could not determine eligibility[a]
Grant KA, McMahon C, Austin MP, Reilly N, Leader L, Ali S. Maternal prenatal anxiety, postnatal caregiving and infants' cortisol responses to the still-face procedure. Dev Psychobiol. 2009;**51**:625-37.	Could not determine eligibility[a]
Grant KA, McMahon C, Reilly N, Austin MP. Maternal sensitivity moderates the impact of prenatal anxiety disorder on infant responses to the still-face procedure. Infant Behav Dev. 2010;**33**:453-62.	Could not determine eligibility[a]
Grigoriadis S, de Camps Meschino D, Barrons E, Bradley L, Eady A, Fishell A, et al. Mood and anxiety disorders in a sample of Canadian perinatal women referred for psychiatric care. Arch Womens Ment Health. 2011;**14**:325-33.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Guedeney N, Fermanian J. Validation study of the French version of the Edinburgh Postnatal Depression Scale (EPDS): new results about use and psychometric properties. Eur Psychiatry. 1998;**13**:83-9.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gutierrez-Zotes A, Labad J, Martin-Santos R, Garcia-Esteve L, Gelabert E, Jover M, et al. Coping strategies and postpartum depressive symptoms: A structural equation modelling approach. Eur Psychiatry. 2015;**30**:701-8.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Gutiérrez -Zotes JA, Farnós A, Vilella E, Labad J. Higher psychoticism as a predictor of thoughts of harming one's infant in postpartum women: a prospective study. Compr Psychiatry. 2013;**54**:1124-9.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Gutiérrez -Zotes A, Labad J, MartinSantos R, GarciaEsteve L, Gelabert E, Jover M, et al. Coping strategies and postpartum depressive symptoms: A structural equation modelling approach. Eur Psychiatry. 2015;**30**:701-8.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Hamdan A, Tamim H. Psychosocial risk and protective factors for postpartum depression in the United Arab Emirates. Arch Womens Ment Health. 2011;**14**:125-33.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Hamdan A, Tamim H. The relationship between postpartum depression and breastfeeding. Int J Psychiatry Med. 2012;**43**:243-59.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Hanusa BH, Scholle SH, Haskett RF, Spadaro K, Wisner KL. Screening for depression in the postpartum period: a comparison of three instruments. J Womens Health. 2008;**17**:585-96.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Harris B, Huckle P, Thomas R, Johns S, Fung H. The use of rating scales to identify post-natal depression. Br J Psychiatry. 1989;**154**:813-7.	Could not determine eligibility[a]
Harris B, Othman S, Davies JA, Weppner GJ, Richards CJ, Newcombe RG, et al. Association between postpartum thyroid dysfunction and thyroid antibodies and depression. BMJ. 1992;**305**:152-6.	No validated interview to assess major depression
Harvey ST, Pun PK. Analysis of positive Edinburgh depression scale referrals to a consultation liaison psychiatry service in a two-year period. Int J Ment Health Nurs. 2007;**16**:161-7.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Hatton DC, HarrisonHohner J, Matarazzo J, Edwards P, Lewy A, Davis L. Missed antenatal depression among high risk women: A secondary analysis. Arch Womens Ment Health. 2007;**10**:121-3.	No validated interview to assess major depression
Henshaw C, Foreman D, Cox J. Postnatal blues: a risk factor for postnatal depression. J Psychosom Obstet Gynecol. 2004;**25**:267-72.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Herz E, Thoma M, Umek W, Gruber K, Linzmayer L, Walcher W, et al. Non-psychotic post-partum depression. Geburtshilfe Frauenheilkd. 1997;**57**:282-8.	Major depression not assessed
Holden JM. Postnatal depression: its nature, effects, and identification using the Edinburgh Postnatal Depression scale. Birth. 1991;**18**:211-21.	No original data
Holt WJ. The detection of postnatal depression in general practice using the Edinburgh postnatal depression scale. N Z M J. 1995;**108**:57.	> 2 weeks between EPDS and diagnostic interview
Howard LM, Flach C, Mehay A, Sharp D, Tylee A. The prevalence of suicidal ideation identified by the Edinburgh Postnatal Depression Scale in postpartum women in primary care: findings from the RESPOND trial. BMC Pregnancy Childbirth. 2011;**11**:57-59.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Huang J, Zhang L, He M, Qiang X, Xiao X, Huang S, et al. Comprehensive evaluation of postpartum depression and correlations between postpartum depression and serum levels of homocysteine in Chinese women. Zhong Nan Da Xue Xue Bao Yi Xue BanZhong Nan Da Xue Xue Bao Yi Xue Ban. 2015;**40**:311-6.	No validated interview to assess major depression
Huang YC, Mathers NJ. Postnatal depression and the experience of South Asian marriage migrant women in Taiwan: survey and semi-structured interview study. Int J Nurs Stud. 2008;**45**:924-31.	Major depression not assessed

Husain N, Cruickshank K, Husain M, Khan S, Tomenson B, Rahman A. Social stress and depression during pregnancy and in the postnatal period in British Pakistani mothers: a cohort study. J Affect Disord. 2012;**140**:268-76.	Could not determine eligibility[a]
Husain N, Kiran T, Sumra A, Naeem Zafar S, Ur Rahman R, Jafri F, et al. Detecting maternal depression in a low-income country: comparison of the self-reporting questionnaire and the Edinburgh Postnatal Depression Scale. J Trop Pediatr. 2014;**60**:129-33.	Could not determine eligibility[a]
Ibanez G, Bernard JY, Rondet C, Peyre H, Forhan A, Kaminski M, et al. Effects of antenatal maternal depression and anxiety on children's early cognitive development: A prospective cohort study. PLOS ONE. 2015;**10**:Art e0135849.	Major depression not assessed
Ikeda M, Hayashi M, Kamibeppu K. The relationship between attachment style and postpartum depression. Attach Hum Dev. 2014;**16**:557-72.	> 2 weeks between EPDS and diagnostic interview
Inglis AJ, Hippman CL, Carrion PB, Honer WG, Austin JC. Mania and depression in the perinatal period among women with a history of major depressive disorders. Arch Womens Ment Health. 2014;**17**:137-43.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Jadresic E, Araya R, Jara C. Validation of the Edinburgh Postnatal Depression Scale (EPDS) in Chilean postpartum women. J Psychosom Obstet Gynecol. 1995;**16**:187-91.	No validated interview to assess major depression
Jaju S, Al Kharusi L, Gowri V. Antenatal prevalence of fear associated with childbirth and depressed mood in primigravid women. Indian J Psychiatry. 2015;**57**:158-61.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Jardri R, Maron M, Pelta J, Thomas P, Codaccioni X, Goudemand M, Delion P. Impact of midwives' training on postnatal depression screening in the first week post delivery: a quality improvement report. Midwifery. 2010;**26**:622-9.	> 2 weeks between EPDS and diagnostic interview

Ji S, Long Q, Newport DJ, Na H, Knight B, Zach EB, et al. Validity of depression rating scales during pregnancy and the postpartum period: impact of trimester and parity. J Psychiatr Res. 2011;**45**:213-9.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Josefsson A, Larsson C, Sydsjö G, Nylander PO. Temperament and character in women with postpartum depression. Arch Womens Ment Health. 2007;**10**:3-7.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Keshavarzi F, Yazdchi K, Rahimi M, Rezaei M, Farnia V, Davarinejad O, et al. Post partum depression and thyroid function. Iran J Psychiatry. 2011;**6**:117-20.	Major depression not assessed
Kirkan TS, Aydin N, Yazici E, Akcali Aslan P, Acemoglu H, Daloglu AG. The depression in women in pregnancy and postpartum period: A follow-up study. Int J Soc Psychiatry. 2015;**61**:343-9.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Klier CM, Muzik M, Dervic K, Mossaheb N, Benesch T, Ulm B, Zeller M. The role of estrogen and progesterone in depression after birth. J Psychiatr Res. 2007;**41**:273-9.	> 2 weeks between EPDS and diagnostic interview
Knorring LV. Book review of Depression in women with focus on the postpartum period. Nord J Psychiatry. 2003;**57**:390.	No validated interview to assess major depression
Kohlhoff J, Hickinbotham R, Knox C, Roach V, Barnett Am B. Antenatal psychosocial assessment and depression screening in a private hospital. Aust N Z J Obstet Gynaecol. 2016;**56**:173-8.	Major depression not assessed
Koss J, Bidzan M, Smutek J, Bidzan L. Influence of perinatal depression on labor-associated fear and mmotional attachment to the child in high-risk pregnancies and the first days after delivery. Med Sci Monit. 2016;**22**:1028-37.	Major depression not assessed
Lai BP, Tang AK, Lee DT, Yip AS, Chung TK. Detecting postnatal depression in Chinese men: a comparison of three instruments. Psychiatry Res. 2010;**180**:80-5.	No pregnant or postpartum women

Lau Y, Wang Y, Yin L, Chan KS, Guo X. Validation of the Mainland Chinese version of the Edinburgh Postnatal Depression Scale in Chengdu mothers. Int J Nurs Stud. 2010;**47**:1139-51.	Could not determine eligibility[a]
Lawrie TA, Hofmeyr GJ, de Jager M, Berk M. Validation of the Edinburgh Postnatal Depression Scale on a cohort of South African women. S Afr Med J. 1998;**88**:1340-4.	No validated interview to assess major depression
Lee DT, Wong CK, Ungvari GS, Cheung LP, Haines CJ, Chung TK. Screening psychiatric morbidity after miscarriage: application of the 30-item General Health Questionnaire and the Edinburgh Postnatal Depression Scale. Psychosom Med. 1997;**59**:207-10.	No pregnant or postpartum women
Lee DT, Yip AS, Chan SS, Tsui MH, Wong WS, Chung TK. Postdelivery screening for postpartum depression. Psychosom Med. 2003;**65**:357-61.	Major depression not assessed
Lee DT, Yip AS, Chiu HF, Chung TK. Screening for postnatal depression using the double-test strategy. Psychosom Med. 2000;**62**:258-63.	Major depression not assessed
Lee DT, Yip AS, Chiu HF, Leung TY, Chung TK. Screening for postnatal depression: are specific instruments mandatory? J Affect Disord. 2001;**63**:233-8.	Major depression not assessed
Lee DT, Yip SK, Chiu HF, Leung TY, Chan KP, Chau IO, et al. Detecting postnatal depression in Chinese women. Validation of the Chinese version of the Edinburgh Postnatal Depression Scale. Br J Psychiatry. 1998;**172**:433-7.	No validated interview to assess major depression
Leverton TJ, Elliott SA. Is the EPDS a magic wand?: 1. A comparison of the Edinburgh Postnatal Depression Scale and health visitor report as predictors of diagnosis on the Present State Examination. J Reprod Infant Psychol. 2000;**18**:279-96.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Lewis BA, Gjerdingen DK, Avery MD, Guo H, Sirard JR, Bonikowske AR, Marcus BH. Examination of a telephone-based exercise intervention for the prevention of	Major depression not assessed

postpartum depression: design, methodology, and baseline data from The Healthy Mom study. Contemp Clin Trials. 2012;**33**:1150-8.	
Lewis BA, Gjerdingen DK, Avery MD, Sirard JR, Guo H, Schuver K, Marcus BH. A randomized trial examining a physical activity intervention for the prevention of postpartum depression: the healthy mom trial. Ment Health Phys Act. 2014;*7*:42-9.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Logsdon MC, Myers JA. Comparative performance of two depression screening instruments in adolescent mothers. J Womens Health. 2010;**19**:1123-8.	Not a sample of adults
ukasik A, B aszczyk K, Wojcieszyn M, Belowska A. Characteristic of affective disorders of the first week of puerperium. Ginekol Pol. 2003;**74**:1194-9.	No validated interview to assess major depression
Lundh W, Gyllang C. Use of the Edinburgh Postnatal Depression Scale in some Swedish child health care centres. Scand J Caring Sci. 1993;**7**:149-54.	No validated interview to assess major depression
Lydsdottir LB, Howard LM, Olafsdottir H, Thome M, Tyrfingsson P, Sigurdsson JF. The mental health characteristics of pregnant women with depressive symptoms identified by the Edinburgh Postnatal Depression Scale. J Clin Psychiatry. 2014;**75**:393-8.	> 2 weeks between EPDS and diagnostic interview
Mallett P, Andrew M, Hunter C, Smith J, Richards C, Othman S, et al. Cognitive function, thyroid status and postpartum depression. Acta Psychiatr Scand. 1995;**91**:243-6.	No validated interview to assess major depression
Maloney DM. Postnatal depression: a study of mothers in the metropolitan area of Perth, Western Australia. Aust J Midwifery. 1998;**11**:18-23.	Major depression not assessed
Mao HJ, Li HJ, Chiu H, Chan WC, Chen SL. Effectiveness of antenatal emotional self-management training program in prevention of postnatal depression in Chinese women. Perspect Psychiatr Care. 2012;**48**:218-24.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Martin-Santos R, Gelabert E, Subira S, Gutierrezzotes A, Langorh K, Jover M, et al. Research Letter: Is neuroticism a risk factor for postpartum depression? Psychol Med. 2012;**42**:1559-65.	No original data
Mason L, Poole H. Healthcare professionals' views of screening for postnatal depression. Community Pract. 2008;**81**:30-4.	No pregnant or postpartum women
Matijasevich A, Munhoz TN, Tavares BF, Barbosa AP, da Silva DM, Abitante MS, et al. Validation of the Edinburgh Postnatal Depression Scale (EPDS) for screening of major depressive episode among adults from the general population. BMC Psychiatry. 2014;**14**:284.	No pregnant or postpartum women
Matthey S, Valenti B, Souter K, Ross-Hamid C. Comparison of four self-report measures and a generic mood question to screen for anxiety during pregnancy in English-speaking women J Affect Disord. 2013;**148**:347-51.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Matthey S. Differentiating between Transient and Enduring distress on the Edinburgh Depression Scale within screening contexts. J Affect Disord. 2016;**196**:252-58.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Matthey S. Using the Edinburgh Postnatal Depression Scale to screen for anxiety disorders. Depress Anxiety. 2008;**25**:926-31.	No pregnant or postpartum women
Mauri M, Banti S, Borri C, Rambelli C, Ramacciotti D, Oppo A, et al. Depressive Symptomatology in Pregnancy Detected with EPDS: the Problem of False Positive. Eur Psychiatry. 2010;**25**:1403.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Mazhari S, Nakhaee N. Validation of the Edinburgh Postnatal Depression Scale in an Iranian sample. Arch Womens Ment Health. 2007;**10**:293-7.	No validated interview to assess major depression

Mazzeo SE, SlofOp't Landt MC, Jones I, Mitchell K, Kendler KS, Neale MC, et al. Associations among postpartum depression, eating disorders, and perfectionism in a population-based sample of adult women. Int J Eat Disord. 2006;**39**:202-11.	Major depression not assessed
McMahon CA, Boivin J, Gibson FL, Hammarberg K, Wynter K, Fisher JR. Older maternal age and major depressive episodes in the first two years after birth: Findings from the Parental Age and Transition to Parenthood Australia (PATPA) study. J Affect Disord. 2015;**175**:454-62.	Major depression not assessed
Meltzer-Brody S, Zerwas S, Leserman J, Von Holle A, Regis T, Bulik C. Eating disorders and trauma history in women with perinatal depression. J Womens Health. 2011;**20**:863-70.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Meuti V, Aceti F, Giacchetti N, Carluccio GM, Zaccagni M, Marini I, et al. Perinatal depression and patterns of attachment: a critical risk factor? Depress Res Treat. 2015;**2015**:105012.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Milgrom J, Gemmill AW, Ericksen J, Burrows G, Buist A, Reece J. Treatment of postnatal depression with cognitive behavioural therapy, sertraline and combination therapy: A randomised controlled trial. Aust N Z J Psychiatry. 2015;**49**:236-245.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Miller L, Gur M, Shanok A, Weissman M. Interpersonal psychotherapy with pregnant adolescents: two pilot studies. J Child Psychol Psychiatry. 2008;**49**:733-42.	Not a sample of adults
Moayedoddin A, Moser D, Nanzer N. The impact of brief psychotherapy centred on parenthood on the anxio-depressive symptoms of mothers during the perinatal period. Swiss Med Wkly. 2013;**143**:w13769.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Murray D, Cox JL, Chapman G, Jones P. Childbirth: life event or start of a long-term difficulty? Further data from the Stoke-on-Trent controlled study of postnatal depression. Br J Psychiatry. 1995;**166**:595-600.	No validated interview to assess major depression
Murray D, Cox JL. Screening for depression during pregnancy with the Edinburgh Depression Scale (EPDS). J Reprod Infant Psychol. 1990;**8**:99-107.	No validated interview to assess major depression
Murray L, Carothers AD. The validation of the Edinburgh Post-natal Depression Scale on a community sample. Br J Psychiatry. 1990;**157**:288-90.	No validated interview to assess major depression
O'Mahen H, Himle JA, Fedock G, Henshaw E, Flynn H. A pilot randomized controlled trial of cognitive behavioral therapy for perinatal depression adapted for women with low incomes. Depress Anxiety. 2013;**30**:679-87.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
O'Neill T. Postnatal depression--aetiological factors. Ir Med J. 1990;**83**:17-18.	> 2 weeks between EPDS and diagnostic interview
Ortiz Collado MA, Saez M, Favrod J, Hatem M. Antenatal psychosomatic programming to reduce postpartum depression risk and improve childbirth outcomes: a randomized controlled trial in Spain and France. BMC Pregnancy & Childbirth. 2014;**14**:22.	Major depression not assessed
Owoeye AO, Aina OF, Morakinyo O. Risk factors of postpartum depression and EPDS scores in a group of Nigerian women. Trop Doct. 2006;**36**:100-3.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Parker G, Hegarty B, Granville-Smith I, Ho J, Paterson A, Gokiert A, Hadzi-Pavlovic D. Is essential fatty acid status in late pregnancy predictive of post-natal depression?. Acta Psychiatr Scand. 2015;**131**:148-56.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Parker GB, Hegarty B, Paterson A, Hadzi-Pavlovic D, Granville-Smith I, Gokiert A. Predictors of post-natal depression are shaped distinctly by the measure of 'depression'. J Affect Disord. 2015;**173**:239-44.	Sample selected for known distress, mental health diagnosis, or psychiatric setting

Patton GC, Romaniuk H, Spry E, Coffey C, Olsson C, Doyle LW, et al. Prediction of perinatal depression from adolescence and before conception (VIHCS): 20-year prospective cohort study. Lancet. 2015;**386**:875-83.	Major depression not assessed
Peindl KS, Wisner KL, Hanusa BH. Identifying depression in the first postpartum year: guidelines for office-based screening and referral. J Affect Disord. 2004;**80**:37-44.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Phillips J, Sharpe L, Nemeth D. Maternal psychopathology and outcomes of a residential mother-infant intervention for unsettled infant behaviour. Aust N Z J Psychiatry. 2010;**44**:280-9.	> 2 weeks between EPDS and diagnostic interview
Piacentini D, Leveni D, Primerano G, Cattaneo M, Volpi L, Biffi G, Mirabella F. Prevalence and risk factors of postnatal depression among women attending antenatal courses. Epidemiologia Psichiatr Soc. 2009;**18**:214-20.	> 2 weeks between EPDS and diagnostic interview
Pitanupong J, Liabsuetrakul T, Vittayanont A. Validation of the Thai Edinburgh Postnatal Depression Scale for screening postpartum depression. Psychiatry Res. 2007;**149**:253-9.	No validated interview to assess major depression
Pollock JI, Manaseki-Holland S, Patel V. Detection of depression in women of child-bearing age in non-Western cultures: a comparison of the Edinburgh Postnatal Depression Scale and the Self-Reporting Questionnaire-20 in Mongolia. J Affect Disord. 2006;**92**:267-71.	Not a sample of adults
Reck C, Stehle E, Reinig K, Mundt C. Maternity blues as a predictor of DSM-IV depression and anxiety disorders in the first three months postpartum. J Affect Disord. 2009;**113**:77-87.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Reck C, Struben K, Backenstrass M, Stefenelli U, Reinig K, Fuchs T, et al. Prevalence, onset and comorbidity of postpartum anxiety and depressive disorders. Acta Psychiatr Scand. 2008;**118**:459-68.	> 2 weeks between EPDS and diagnostic interview

Regmi S, Sligl W, Carter D, Grut W, Seear M. A controlled study of postpartum depression among Nepalese women: validation of the Edinburgh Postpartum Depression Scale in Kathmandu. Trop Med Int Health. 2002;**7**:378-82.	Major depression not assessed
Robakis TK, Williams KE, Crowe S, Kenna H, Gannon J, Rasgon NL. Optimistic outlook regarding maternity protects against depressive symptoms postpartum. Arch Womens Ment Health. 2015;**18**:197-208.	No validated interview to assess major depression
Roca A, Imaz ML, Torres A, Plaza A, Subira S, Valdes M, et al. Unplanned pregnancy and discontinuation of SSRIs in pregnant women with previously treated affective disorder. J Affect Disord. 2013;**150**:807-13.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Rojas G, Fritsch R, Solis J, Gonzalez M, Guajardo V, Araya R. Quality of life of women depressed in the post-partum period. Rev Med Chil. 2006;**134**:713-20.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Rubertsson C, Borjesson K, Berglund A, Josefsson A, Sydsjo G. The Swedish validation of Edinburgh Postnatal Depression Scale (EPDS) during pregnancy. Nord J Psychiatry. 2011;**65**:414-8.	No validated interview to assess major depression
Saleh ES, El-Bahei W, El-Hadidy MA, Zayed A. Predictors of postpartum depression in a sample of Egyptian women. Neuropsychiatr Dis Treat. 2012;**9**:15-24.	EPDS not administered
Sanjuan J, MartinSantos R, GarciaEsteve L, Carot JM, Guillamat R, GutierrezZotes A, et al. Mood changes after delivery: Role of the serotonin transporter gene. Br J Psychiatry. 2008;**193**:383-8.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Santos IS, Matijasevich A, Tavares BF, Barros AJ, Botelho IP, Lapolli C, et al. Validation of the Edinburgh Postnatal Depression Scale (EPDS) in a sample of mothers from the 2004 Pelotas Birth Cohort Study.Cad Saude Publica. 2007;**23**:2577-88.	No validated interview to assess major depression

Santos IS, Matijasevich A, Tavares BF, da Cruz Lima AC, Riegel RE, Lopes BC. Comparing validity of Edinburgh scale and SRQ20 in screening for post-partum depression. Clin Pract Epidemiol Ment Health. 2007;**3**:18.	No validated interview to assess major depression
Savarimuthu RJ, Ezhilarasu P, Charles H, Antonisamy B, Kurian S, Jacob KS. Post-partum depression in the community: a qualitative study from rural South India. Int J Soc Psychiatry. 2010;**56**:94-102.	Could not determine eligibility[a]
Séjourné N, Alba J, Onorrus M, Goutaudier N, Chabrol H. Intergenerational transmission of postpartum depression. J Reprod Infant Psychol. 2011;**29**:115-24.	No validated interview to assess major depression
Seth S, Lewis AJ, Saffery R, Lappas M, Galbally M. Maternal prenatal mental health and placental 11 beta-HSD2 gene expression: initial findings from the Mercy Pregnancy and Emotional Wellbeing study. Int J Mol Sci. 2015;**16**:27482-96.	Major depression not assessed
Simpson W, Glazer M, Michalski N, Steiner M, Frey BN. Comparative efficacy of the generalized anxiety disorder 7-item scale and the Edinburgh Postnatal Depression Scale as screening tools for generalized anxiety disorder in pregnancy and the postpartum period. Can J Psychiatry. 2014;**59**:434-40.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Sit DK, Flint C, Svidergol D, White J, Wimer M, Bish B, Wisner KL. Best practices: an emerging best practice model for perinatal depression care. Psychiatr Serv. 2009;**60**:1429-31.	No validated interview to assess major depression
Slade P, Morrell CJ, Rigby A, Ricci K, Spittlehouse J, Brugha TS. Postnatal women's experiences of management of depressive symptoms: a qualitative study. Br J Gen Pract. 2010;**60**:e440-e448.	Major depression not assessed

Smith-Nielsen J, Steele H, Mehlhase H, Cordes K, Steele M, Harder S, Vaever MS. Links among high EPDS scores, state of mind regarding attachment, and symptoms of personality disorder. J Pers Disord. 2015;**29**:771-93.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Sundaram S, Harman JS, Cook RL. Maternal morbidities and postpartum depression: An analysis using the 2007 and 2008 pregnancy risk assessment monitoring system. Womens Health Issues. 2014;**24**:e381-8.	EPDS not administered
Sutter-Dallay AL, Giaconne-Marcesche V, Glatigny-Dallay E, Verdoux H. Women with anxiety disorders during pregnancy are at increased risk of intense postnatal depressive symptoms: a prospective survey of the MATQUID cohort. Eur Psychiatry. 2004;**19**:459-63.	> 2 weeks between EPDS and diagnostic interview
Tam LW, Newton RP, Dern M, Parry BL. Screening women for postpartum depression at well baby visits: resistance encountered and recommendations. Arch Womens Ment Health. 2002;**5**:79-82.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Tan EC, Chua TE, Lee TMY, Tan HS, Ting JLY, Chen HY. Case-control study of glucocorticoid receptor and corticotrophin-releasing hormone receptor gene variants and risk of perinatal depression. BMC Pregnancy Childbirth. 2015;**15**:283.	Major depression not assessed
Tang Y, Shi S, Lu W, Chen Y, Wang Q, Zhu Y, et al. Prenatal psychological prevention trial on postpartum anxiety and depression. Chin Ment Health J. 2009;**23**:83-89.	Could not determine eligibility[a]
Teng HW, Hsu CS, Shih SM, Lu ML, Pan JJ, Shen WW. Screening postpartum depression with the Taiwanese version of the Edinburgh Postnatal Depression scale. Compr Psychiatry. 2005;**46**:261-65.	Could not determine eligibility[a]

Tesfaye M, Hanlon C, Wondimagegn D, Alem A. Detecting postnatal common mental disorders in Addis Ababa, Ethiopia: validation of the Edinburgh Postnatal Depression Scale and Kessler Scales. J Affect Disord. 2010;**122**:102-8.	No validated interview to assess major depression
Tharner A, Luijk MPCM, van IJzendoorn MH, BakermansKranenburg MJ, Jaddoe VWV, Hofman A, et al. Maternal lifetime history of depression and depressive symptoms in the prenatal and early postnatal period do not predict infant-mother attachment quality in a large, population-based Dutch cohort study. Attach Hum Dev. 2012;**14**:63-81.	> 2 weeks between EPDS and diagnostic interview
Thorpe K. A study of the use of the Edinburgh Postnatal Depression Scale with parent groups outside the postpartum period. J Reprod Infant Psychol. 1993;**11**:119-25.	No pregnant or postpartum women
Tietz A, Zietlow AL, Reck C. Maternal bonding in mothers with postpartum anxiety disorder: the crucial role of subclinical depressive symptoms and maternal avoidance behaviour. Arch Womens Ment Health. 2014;**17**:433-42.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Ueda M, Yamashita H, Yoshida K. Impact of infant health problems on postnatal depression: pilot study to evaluate a health visiting system. Psychiatry Clin Neurosci. 2006;**60**:182-9.	> 2 weeks between EPDS and diagnostic interview
Uguz F, Akman C, Sahingoz M, Kaya N, Kucur R. One year follow-up of post-partum-onset depression: the role of depressive symptom severity and personality disorders. J Psychosom Obstet Gynecol. 2009;**30**:141-5.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Uwakwe R, Okonkwo JE. Affective (depressive) morbidity in puerperal Nigerian women: validation of the Edinburgh Postnatal Depression Scale. Acta Psychiatr Scand. 2003;**107**:251-9.	No validated interview to assess major depression

Venkatesh KK, Zlotnick C, Triche EW, Ware C, Phipps MG. Accuracy of brief screening tools for identifying postpartum depression among adolescent mothers. Pediatrics. 2014;**133**:e45-e45.	Not a sample of adults
Venter MD, Smets J, Raes F, Wouters K, Franck E, Hanssens M, et al. Impact of childhood trauma on postpartum depression: A prospective study. Arch Womens Ment Health. 2016;**19**:337-42.	Major depression not assessed
Verkerk GJ, Denollet J, Van Heck GL, Van Son MJ, Pop VJ. Personality factors as determinants of depression in postpartum women: a prospective 1-year follow-up study. Psychosom Med. 2005;**67**:632-7.	No validated interview to assess major depression
Verkerk GJM, Pop VJM, Van Son MJM, Van Heck GL. Prediction of depression in the postpartum period: A longitudinal follow-up study in high-risk and low-risk women. J Affect Disord. 2003;**77**:159-66.	> 2 weeks between EPDS and diagnostic interview
Viktorin A, Meltzer-Brody S, Kuja-Halkola R, Sullivan PF, Landen M, Lichtenstein P, Magnusson PK. Heritability of perinatal depression and genetic overlap with nonperinatal depression. Am J Psychiatry. 2016;**173**:158-65.	EPDS not administered
Wang Y, Guo X, Lau Y, Chan KS, Yin L, Chen J. Psychometric evaluation of the Mainland Chinese version of the Edinburgh Postnatal Depression Scale. Int J Nurs Stud. 2009;**46**:813-23.	Could not determine eligibility[a]
Warner R, Appleby L, Whitton A, Faragher B. Attitudes toward motherhood in postnatal depression: development of the Maternal Attitudes Questionnaire. J Psychosom Res. 1997;**43**:351-8.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Warnock FF, Bakeman R, Shearer K, Misri S, Oberlander T. Caregiving behavior and interactions of prenatally depressed mothers (antidepressant-treated and non-	Could not determine eligibility[a]

antidepressant-treated) during newborn acute pain. Infant Ment Health J. 2009;**30**:384-406.	
Weobong B, Akpalu B, Doku V, Agyei SO, Hurt L, Kirkwood B, Prince M. The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana. J Affect Disord. 2009;**113**:109-17.	No validated interview to assess major depression
Werrett J, Clifford C. Validation of the Punjabi version of the Edinburgh postnatal depression scale (EPDS). Int J Nurs Stud. 2006;**43**:227-36.	Major depression not assessed
Wickberg B, Hwang CP. Counselling of postnatal depression: a controlled study on a population based Swedish sample. J Affect Disord. 1996;**39**:209-16.	No validated interview to assess major depression
Wickberg B, Hwang CP. The Edinburgh Postnatal Depression Scale: validation on a Swedish community sample. Acta Psychiatr Scand. 1996;**94**:181-84.	No validated interview to assess major depression
Wu M, Li X, Feng B, Wu H, Qiu C, Zhang W. Correlation between sleep quality of third-trimester pregnancy and postpartum depression. Med Sci Monit. 2014;**20**:2740-5.	Could not determine eligibility[a]
Yamashita H, Yoshida K, Nakano H, Tashiro N. Postnatal depression in Japanese women. Detecting the early onset of postnatal depression by closely monitoring the postpartum mood. J Affect Disord. 2000;**58**:145-54.	No validated interview to assess major depression
Yonkers KA, Ramin SM, Rush AJ, Navarrete CA, Carmody T, March D, et al. Onset and persistence of postpartum depression in an inner-city maternal health clinic system. Am J Psychiatry. 2001;**158**:1856-63.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Yoshida K, Yamashita H, Ueda M, Tashiro N. Postnatal depression in Japanese mothers and the reconsideration of 'Satogaeri bunben'. Pediatr Int. 2001;**43**:189-93.	No validated interview to assess major depression

Zammit S, Thomas K, Thompson A, Horwood J, Menezes P, Gunnell D, et al. Maternal tobacco, cannabis and alcohol use during pregnancy and risk of adolescent psychotic symptoms in offspring. Br J Psychiatry. 2009;**195**:294-300.	No pregnant or postpartum women
Zelkowitz P, Milet TH. Postpartum psychiatric disorders: Their relationship to psychological adjustment and marital satisfaction in the spouses. J Abnorm Psychol. 1996;**105**:281-5.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Zlotnick C, Capezza NM, Parker D. An interpersonally based intervention for low-income pregnant women with intimate partner violence: A pilot study. Arch Womens Ment Health. 2011;**14**:55-65.	Sample selected for known distress, mental health diagnosis, or psychiatric setting
Zubaran C, Foresti K, Schumacher MV, Amoretti AL, Thorell MR, Muller LC. The correlation between postpartum depression and health status. Matern & Child Health J. 2010;**14**:751-7.	> 2 weeks between EPDS and diagnostic interview

[a]It was not possible to determine eligibility based on published paper, and we were not able to obtain clarification from authors despite multiple attempts

eTable2. Characteristics of included 49 primary studies included in the IPDMA dataset

First author, Year	Country	Recruited population	Diagnostic interview	Classification system	Total N	N (%) Major depression
Aceti, 2012[1]	Italy	Pregnant women in the third trimester	SCID	DSM-IV	44	22 (50)
Alvarado, 2015[2]	Chile	Pregnant women up to 28 weeks gestation	MINI	DSM-IV	111	38 (34)
Alvarado–Esquivel, 2006[3]	Mexico	Women within 3 months postpartum	MINI	DSM-IV	91	10 (11)
Alvarado–Esquivel, 2016[4]	Mexico	Pregnant women recruited at a public hospital	MINI	DSM-IV	184	12 (7)
Bakare, 2014[5]	Nigeria	Postpartum women	MINI	DSM-IV	405	62 (15)
Barnes, 2009[6]	UK	Socially disadvantaged mothers at 2 months postpartum	SCID	DSM-III-R	347	25 (7)
Bavle, 2016[7]	India	Pregnant women recruited from an outpatient obstetrics department in a tertiary care hospital	SCID	DSM-IV	318	6 (2)
Beck, 2001[8]	USA	Postpartum women	SCID	DSM-IV	150	18 (12)
Bunevicius, 2009[9]	Lithuania	Pregnant women 12 to 16 weeks gestation attending an obstetric clinic	SCID	DSM-III-R	230	12 (5)
Castro e Couto, 2015[10]	Brazil	Women in their second trimester of pregnancy recruited at antenatal care in a public hospital	MINI	DSM-IV-TR	173	36 (21)

Chaudron, 2010[11]	USA	Postpartum women recruited from Well-Child Care visits with infants 0-14 months of age	SCID	DSM-IV	187	70 (37)
Comasco, 2016[12]	Sweden	Pregnant women	MINI	DSM-IV	220	18 (8)
de Figueiredo, 2015[13]	Brazil	Postpartum women enrolled in prenatal care outpatient services	SCID	DSM-IV	241	94 (39)
Eapen, 2013[14]	Australia	Women attending an antenatal clinic	MINI	DSM-IV	131	26 (20)
Felice, 2004[15]	Malta	Pregnant women attending an antenatal clinic	CIS-R	ICD-10	226	32(14)
Fernandes, 2011[16]	India	Rural women in their third trimester	MINI	DSM-IV	133	27 (20)
Figueira, 2009[17]	Brazil	Postpartum women recruited from hospitalization records	MINI	DSM-IV	239	18 (8)
Fisher, 2010[18]	Australia	Postpartum women recruited in Australian maternal and child health centres at 6 months postpartum	CIDI	DSM-IV	192	1 (1)
Garcia–Esteve, 2003[19]	Spain	Women at 6 weeks postpartum	SCID	DSM-III-R	334	36 (11)
Giardinelli, 2012[20]	Italy	Women between 28 and 32 weeks gestation recruited from a obstetric course	SCID	DSM-IV	588	28 (5)
Helle, 2015[21]	Germany	Women with very low birthweight and normal weight infants between 4 and 6 weeks	SCID	DSM-IV	224	12 (5)
Hickey, 1997[22]	Australia	Postpartum women recruited in the hospital after delivery	SCID	DSM-III-R	72	31 (43)

Howard, 2018[23]	UK	Pregnant women recruited from an inner-city London maternity service	SCID	DSM-IV	527	130 (25)
Imbula, 2012[24]	Democratic Republic of Congo	Women between 1 and 10 months postpartum recruited from 'well-baby' clinics	MINI	DSM-IV-TR	117	29 (25)
Khalifa, 2015[25]	Sudan	Women at 3 months postpartum	MINI	ICD-10	40	18 (45)
Leonardou, 2009[26]	Greece	Postpartum women recruited from private and public maternity wards on their second day postpartum	SCID	DSM-III-R	81	4 (5)
Navarro, 2007[27]	Spain	Women presenting for postpartum care at 6 weeks	SCID	DSM-IV	401	84 (21)
Nakić Radoš, 2013[28]	Croatia	Women between 6 and 8 weeks postpartum	SCID	DSM-IV-TR	272	10 (4)
Pawlby, 2008[29]	UK	Postpartum women at 3 months	CIS	ICD-9	190	34 (18)
Phillips, 2009[30]	Australia	Postpartum women with unsettled infants	SCID	DSM-IV	158	42 (27)
Prenoveau, 2013[31]	UK	Postpartum women at 10 months recruited from mixed health centres.	SCID	DSM-IV	219	20 (9)
Robertson–Blackmore, 2013[32]	USA	Women at 18 weeks gestation	SCID	DSM-IV-TR	358	29 (8)
Rochat, 2013[33]	South Africa	Women recruited from their antenatal appointment at a primary health care clinic between 26 and 34 weeks of pregnancy	SCID	DSM-IV	104	50 (48)

Roomruangwong, 2016[34]	Thailand	Pregnant women at the end of their term	MINI	DSM-IV-TR	126	1 (1)
Rowe, 2008[35]	Australia	English speaking women admitted with their up to 1-year-old infants to private parenting centers	CIDI	DSM-IV	137	25 (18)
Siu, 2012[36]	China	Postpartum women	SCID	DSM-IV	805	126 (16)
Stewart, 2013[37]	Malawi	Pregnant women attending an antenatal clinic	SCID	DSM-IV	186	34 (18)
Su, 2007[38]	Taiwan	Women in their second and third trimesters	MINI	DSM-IV	185	23 (12)
Tandon, 2012[39]	USA	Pregnant and postpartum women enrolled in home visitation programs	SCID	DSM IV	89	25 (28)
Tendais, 2014[40]	Portugal	Pregnant women recruited in an obstetrics outpatient unit	SCID	DSM-IV	141	18 (13)
Thiagayson, 2013[41]	Singapore	Inpatient high-risk pregnant women at 23 or more weeks of gestation	MINI	DSM-IV	200	22 (11)
Tissot, 2015[42]	Switzerland	Women at 3 months postpartum	DIGS	DSM-IV	65	4 (6)
Töreki, 2013[43]	Hungary	Women at 12 weeks gestation	SCID	DSM-IV	219	7 (3)
Töreki, 2014[44]	Hungary	Women between 6 and 8 weeks postpartum	SCID	DSM-IV	265	8 (3)
Tran, 2011[45]	Vietnam	Pregnant and postpartum Vietnamese women recruited from the commune health centre	SCID	DSM-IV	359	52 (14)

Turner, 2009[46]	Italy	Women from a regional epilepsy center in Italy between 5 and 8 weeks postpartum	SCID	DSM-IV-TR	54	5 (9)
Usuda, 2016[47]	Japan	Pregnant women between 12 and 24 weeks of gestation recruited at maternity hospital	MINI	DSM-IV	177	2 (1)
Vega–Dienstmaier, 2002[48]	Peru	Women up to 12 months postpartum	SCID	DSM-IV	306	19 (6)
Yonkers, 2014[49]	USA	Women at 17 weeks gestation	CIDI	DSM-IV	2634	170 (6)

Abbreviations: CIDI: Composite International Diagnostic Interview; CIS-R: Clinical Interview Schedule – Revised; DIS: Diagnostic Interview Schedule; DIGS: Diagnostic Interview for Genetic Studies; DSM: Diagnostic and Statistical Manual of Mental Disorders; EPDS: Edinburgh Postnatal Depression Scale; ICD: International Classification of Diseases; MINI: Mini International Neuropsychiatric Interview; NR: Not Reported; SCID: Structured Clinical Interview for DSM Disorders; UK: United Kingdom; USA: United States of America.

eTable2 References

1. Aceti F, Aveni F, Baglioni V, Carluccio GM, Colosimo D, Giacchetti N, et al. Perinatal and postpartum depression: from attachment to personality. A pilot study. J Psychopathology. 2012;18:328-34.
2. Alvarado R, Jadresic E, Guajardo V, Rojas G. First validation of a Spanish-translated version of the Edinburgh postnatal depression scale (EPDS) for use in pregnant women. A Chilean study. Arch Womens Ment Health. 2015;18:607-12.
3. Alvarado-Esquivel C, Sifuentes-Alvarez A, Salas-Martinez C, Martínez-García S. Validation of the Edinburgh Postpartum Depression Scale in a population of puerperal women in Mexico. Clin Pract Epidemiol Ment Health. 2006;2:33.
4. Alvarado-Esquivel C, Sifuentes-Alvarez A, Salas-Martinez C. Detection of mental disorders other than depression with the Edinburgh Postnatal Depression Scale in a sample of pregnant women in northern Mexico. Mental Illness. 2016;8:6021.
5. Bakare MO, Okoye JO, Obindo JT. Introducing depression and developmental screenings into the National Programme on Immunization (NPI) in southeast Nigeria: an experimental cross-sectional assessment. Gen Hosp Psychiatry. 2014;36:105-12.
6. Barnes J, Senior R, MacPherson K. The utility of volunteer home□visiting support to prevent maternal depression in the first year of life. Child Care Health Dev. 2009;35:807-16.
7. Bavle AD, Chandahalli AS, Phatak AS, Rangaiah N, Kuthandahalli SM, Nagendra PN. Antenatal depression in a tertiary care hospital. Indian J Psychol Med. 2016;38:31.
8. Beck CT, Gable RK. Comparative analysis of the performance of the Postpartum Depression Screening Scale with two other depression instruments. Nurs Res. 2001;50:242-50.
9. Bunevicius A, Kusminskas L, Pop VJ, Pedersen CA, Bunevicius R. Screening for antenatal depression with the Edinburgh Depression Scale. J Psychosom Obstet Gynecol. 2009;30:238-43.
10. Castro E Couto T, Martins Brancaglion MY, Nogueira Cardoso M, Bergo Protzner A, Duarte Garcia F, Nicolato R, Lopes P Aguiar RA, Vitor Leite H, Corrêa H. What is the best tool for screening antenatal depression? J Affect Disord. 2015;178:12-7.

11. Chaudron LH, Szilagyi PG, Tang W, Anson E, Talbot NL, Wadkins HI, et al. Accuracy of depression screening tools for identifying postpartum depression among urban mothers. Pediatrics. 2010:125:e609-17.
12. Comasco E, Gulinello M, Hellgren C, Skalkidou A, Sylven S, Sundström-Poromaa I. Sleep duration, depression, and oxytocinergic genotype influence prepulse inhibition of the startle reflex in postpartum women. Eur Neuropsychopharmacol. 2016;26:767-76.
13. de Figueiredo FP, Parada AP, Cardoso VC, Batista RF, da Silva AA, Barbieri MA, et al. Postpartum depression screening by telephone: a good alternative for public health and research. Arch Womens Ment Health. 2015;18:547-53.
14. Eapen V, Johnston D, Apler A, Rees S, Silove DM. Adult separation anxiety during pregnancy and its relationship to depression and anxiety. J Perinat Med. 2013;41:159-63.
15. Felice E, Saliba J, Grech V, Cox J. Prevalence rates and psychosocial characteristics associated with depression in pregnancy and postpartum in Maltese women. Journal of Affective Disorders.2004;82:297-301.
16. Fernandes MC, Srinivasan K, Stein AL, Menezes G, Sumithra RS, Ramchandani PG. Assessing prenatal depression in the rural developing world: a comparison of two screening measures. Arch Womens Ment Health. 2011;14:209-16.
17. Figueira P, Corrêa H, Malloy-Diniz L, Romano-Silva MA. Edinburgh Postnatal Depression Scale for screening in the public health system. Rev Saude Publica. 2009;43:79-84.
18. Fisher JR, Wynter KH, Rowe HJ. Innovative psycho-educational program to prevent common postpartum mental disorders in primiparous women: a before and after controlled study. BMC Public Health. 2010;10:432.
19. Garcia-Esteve L, Ascaso C, Ojuel J, Navarro P. Validation of the Edinburgh postnatal depression scale (EPDS) in Spanish mothers. J Affect Disord. 2003;75:71-6.
20. Giardinelli L, Innocenti A, Benni L, Stefanini MC, Lino G, Lunardi C, et al. Depression and anxiety in perinatal period: prevalence and risk factors in an Italian sample. Arch Womens Ment Health. 2012;15:21-30.
21. Helle N, Barkmann C, Bartz-Seel J, Diehl T, Ehrhardt S, Hendel A, et al. Very low birth-weight as a risk factor for postpartum depression four to six weeks postbirth in mothers

and fathers: Cross-sectional results from a controlled multicentre cohort study. J Affect Disord. 2015;180:154-61.

22. Hickey AR, Boyce PM, Ellwood D, Morris-Yates AD. Early discharge and risk for postnatal depression. Med J Aust. 1997;167:244-7.
23. Howard LM, Ryan EG, Trevillion K, Anderson F, Bick D, Bye A, et al. Accuracy of the Whooley questions and the Edinburgh Postnatal Depression Scale in identifying depression and other mental disorders in early pregnancy. Br J Psychiatry. 2018;212:50-6.
24. Imbula BE, Okitundu EL, Mampunza SM. Postpartum depression in Kinshasa (DR Congo): prevalence and risk factors. Med Sante Trop. 2012;22:379-84.
25. Khalifa DS, Glavin K, Bjertness E, Lien L. Postnatal depression among Sudanese women: prevalence and validation of the Edinburgh Postnatal Depression Scale at 3 months postpartum. Health Care Women Int. 2015;7:677-84.
26. Leonardou AA, Zervas YM, Papageorgiou CC, Marks MN, Tsartsara EC, Antsaklis A, et al. Validation of the Edinburgh Postnatal Depression Scale and prevalence of postnatal depression at two months postpartum in a sample of Greek mothers. J Reprod Infant Psychol. 2009;27:28-39.
27. Navarro P, Ascaso C, Garcia-Esteve L, Aguado J, Torres A, Martín-Santos R. Postnatal psychiatric morbidity: a validation study of the GHQ-12 and the EPDS as screening tools. Gen Hosp Psychiatry. 2007;29:1-7.
28. Naki Radoš, Tadinac M, Herman R. Validation study of the Croatian version of the Edinburgh Postnatal Depression Scale (EPDS). Suvrem Psihol. 2013;16:203-18.
29. Pawlby S, Sharp D, Hay D, O'Keane V. Postnatal depression and child outcome at 11 years: the importance of accurate diagnosis. Journal of Affective Disorders. 2008;107:241-5.
30. Phillips J, Charles M, Sharpe L, Matthey S. Validation of the subscales of the Edinburgh Postnatal Depression Scale in a sample of women with unsettled infants. J Affect Disord. 2009;118:101-12.

31. Prenoveau J, Craske M, Counsell N, West V, Davies B, Cooper P, et al. Postpartum GAD is a risk factor for postpartum MDD: the course and longitudinal relationships of postpartum GAD and MDD. Depress Anxiety. 2013;30:506-14.
32. Robertson-Blackmore E, Putnam FW, Rubinow DR, Matthieu M, Hunn JE, Putnam KT, Moynihan JA, O'Connor TG. Antecedent trauma exposure and risk of depression in the perinatal period. J Clin Psychiatry. 2013;74:e942-8.
33. Rochat TJ, Tomlinson M, Newell ML, Stein A. Detection of antenatal depression in rural HIV-affected populations with short and ultrashort versions of the Edinburgh Postnatal Depression Scale (EPDS). Arch Womens Ment Health. 2013;16:401-10.
34. Roomruangwong C, Kanchanatawan B, Sirivichayakul S, Maes M. Antenatal depression and hematocrit levels as predictors of postpartum depression and anxiety symptoms. Psychiatry Res. 2016;238:211-7.
35. Rowe HJ, Fisher JR, Loh WM. The Edinburgh Postnatal Depression Scale detects but does not distinguish anxiety disorders from depression in mothers of infants. Arch Womens Ment Health. 2008;11:103-8.
36. Siu BW, Leung SS, Ip P, Hung SF, O'Hara MW. Antenatal risk factors for postnatal depression: a prospective study of Chinese women at maternal and child health centres. BMC Psychiatry. 2012;12:22.
37. Stewart RC, Umar E, Tomenson B, Creed F. Validation of screening tools for antenatal depression in Malawi—A comparison of the Edinburgh Postnatal Depression Scale and Self Reporting Questionnaire. J Affect Disord. 2013;150:1041-7.
38. Su KP, Chiu TH, Huang CL, Ho M, Lee CC, Wu PL, et al. Different cutoff points for different trimesters? The use of Edinburgh Postnatal Depression Scale and Beck Depression Inventory to screen for depression in pregnant Taiwanese women. Gen Hosp Psychiatry. 2007;29:436-41.
39. Tandon SD, Cluxton-Keller F, Leis J, Le HN, Perry DF. A comparison of three screening tools to identify perinatal depression among low-income African American women. J Affect Disord. 2012;136:155-62.

40. Tendais I, Costa R, Conde A, Figueiredo B. Screening for depression and anxiety disorders from pregnancy to postpartum with the EPDS and STAI. Span J of Psychol. 2014;17:E7.
41. Thiagayson P, Krishnaswamy G, Lim ML, Sung SC, Haley CL, Fung DS, et al. Depression and anxiety in Singaporean high-risk pregnancies—prevalence and screening. Gen Hosp Psychiatry. 2013;35:112-6.
42. Tissot H, Favez N, Frascarolo-Moutinot F, Despland JN. Assessing postpartum depression: Evidences for the need of multiple methods. Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology. 2015;65:61-6.
43. Töreki A, Andó B, Keresztúri A, Sikovanyecz J, Dudas RB, Janka Z, et al. The Edinburgh Postnatal Depression Scale: translation and antepartum validation for a Hungarian sample. Midwifery. 2013;29:308-15.
44. Töreki A, Andó B, Dudas RB, Dweik D, Janka Z, Kozinszky Z, Keresztúri A. Validation of the Edinburgh Postnatal Depression Scale as a screening tool for postpartum depression in a clinical sample in Hungary. Midwifery. 2014;30:911-8.
45. Tran TD, Tran T, La B, Lee D, Rosenthal D, Fisher J. Screening for perinatal common mental disorders in women in the north of Vietnam: a comparison of three psychometric instruments. J Affect Disord. 2011;133:281-93.
46. Turner K, Piazzini A, Franza A, Marconi AM, Canger R, Canevini MP. Epilepsy and postpartum depression. Epilepsia. 2009;50:24-7.
47. Usuda K, Nishi D, Makino M, Tachimori H, Matsuoka Y, Sano Y, et al. Prevalence and related factors of common mental disorders during pregnancy in Japan: a cross-sectional study. Biopsychosoc Med. 2016;10:17.
48. Vega-Dienstmaier JM, Mazzotti GS, Campos MS. Validation of a Spanish version of the Edinburgh postnatal depression scale. Actas Esp Psiquiatr. 2002;30:106-11.
49. Yonkers KA, Smith MV, Forray A, Epperson CN, Costello D, Lin H, Belanger K. Pregnant women with posttraumatic stress disorder and risk of preterm birth. JAMA Psychiatry. 2014;71:897-904.

eTable3. Characteristics of eligible primary studies that did not provide data for the present study (N = 24)

First author, Year	Country	Recruited population	Diagnostic interview	Total N	N (%) Major depression
Adewuya, 2006[1]	Nigeria	Pregnancy, 32-36 weeks	MINI	86	9 (10)
Adouard, 2005[2]	France	Pregnancy, 28-34 weeks	MINI	60	15 (25)
Agoub, 2005[3]	Morocco	Postpartum, 2-3 weeks	MINI	144	27 (19)
Aydin, 2004[4]	Turkey	Postpartum, within first year	SCID	341	28 (8)
Banti, 2011[5]	Italy	Pregnancy, 3 months	SCID	1066	NR
Barnett, 1999[6]	Australia	Postpartum, 6 weeks	DIS	316	21 (7)
Benvenuti, 1999[7]	Italy	Postpartum, 8-12 weeks	MINI	113	18 (16)
Bergink, 2011[8]	The Netherlands	Pregnancy, 12 weeks	CIDI	845	47 (6)
Berle, 2003[9]	Norway	Postpartum, 6-12 weeks	MINI	100	27 (27)
Brodey, 2016[10]	USA	Pregnant/Postpartum mixed sample. Postpartum sample was 0-150 days	SCID	879	NR

Chibanda, 2010[11]	Zimbabwe	Postpartum, 6 weeks	SCID	210	NR
Christl, 2013[12]	Australia	Postpartum, between 0-12 weeks	MINI	232	13 (6)
Crotty, 2004[13]	Ireland	Postpartum, 6 weeks	SCAN	113	NR
Gausia, 2007[14]	Bangladesh	Postpartum, 6-8 weeks	SCID	100	3 (3)
Gorman, 2004[15]	France, Ireland, Italy, USA, UK, Portugal, Austria, Switzerland	Pregnancy, second or third semester	SCID	289	10 (4)
Li, 2011[16]	China	Postpartum, 2-12 weeks	SCID	387	24 (6)
Mahmud, 2003[17]	Malaysia	Postpartum, 4-12 weeks	CIDI	64	9 (14)
Matthey, 2001[18]	Australia	Postpartum, 6-7 weeks	DIS	230	11 (5)
Moses–Kolko, 2012[19]	USA	Postpartum, 0-16 weeks	SCID	33	13 (39)
O'Brien, 2004[20]	UK	Postpartum, ≤2 years	CIS-R	216	31 (14)
Pedersen, 2016[21]	USA	Pregnancy, 35-36 weeks	MINI	199	NR

Pinheiro 2013[22]	Brazil	Postpartum, 45-90 days	MINI	207	27 (13)
Priest, 2003[23]	Australia	Postpartum, 2 months	SADS	292	NR
Stuebe, 2013[24]	USA	Pregnancy, 3rd trimester	SCID	47	8 (17)

Abbreviations: CIDI: Composite International Diagnostic Interview; CIS-R: Clinical Interview Schedule – Revised; DIS: Diagnostic Interview Schedule; MINI: Mini International Neuropsychiatric Interview; NR: Not Reported; SADS: Schedule for Affective Disorders and Schizophrenia; SCAN: Schedule for Clinical Assessment in Neuropsychiatry; SCID: Structured Clinical Interview for DSM Disorders; UK: United Kingdom; USA: United States of America.

eTable3 References

1. Adewuya AO, Ola BA, Dada AO, Fasoto OO. Validation of the Edinburgh Postnatal Depression Scale as a screening tool for depression in late pregnancy among Nigerian women. J Psychosom Obstet Gynecol. 2006;27:267-72.
2. Adouard F, Glangeaud-Freudenthal NM, Golse B. Validation of the Edinburgh Postnatal Depression Scale (EPDS) in a sample of women with high-risk pregnancies in France. Arch Womens Ment Health. 2005;8:89-95.
3. Agoub M, Moussaoui D, Battas O. Prevalence of postpartum depression in a Moroccan sample. Arch Womens Ment Health. 2005;8:37-43.
4. Aydin N, Inandi T, Yigit A, Hodoglugil NN. Validation of the Turkish version of the Edinburgh Postnatal Depression Scale among women within their first postpartum year. Soc Psychiatry Psychiatr Epidemiol. 2004;39:483-6.
5. Banti S, Mauri M, Oppo A, Borri C, Rambelli C, Ramacciotti D, et al. From the third month of pregnancy to 1 year postpartum. Prevalence, incidence, recurrence, and new onset of depression. Results from the perinatal depression-research & screening unit study. Compr Psychiatry. 2011;52:343-51.
6. Barnett B, Matthey S, Gyaneshwar R. Screening for postnatal depression in women of non-English speaking background. Archives of Women's Mental Health. 1999;2:67.
7. Benvenuti P, Ferrara M, Niccolai C, Valoriani V, Cox JL. The Edinburgh Postnatal Depression Scale: validation for an Italian sample. J Affective Disord. 1999;53:137-41.
8. Bergink V, Kooistra L, Lambregtse-van den Berg MP, Wijnen H, Bunevicius R, van Baar A, Pop V. Validation of the Edinburgh Depression Scale during pregnancy. J Psychosom Res. 2011;70:385-9.
9. Berle JØ, Aarre TF, Mykletun A, Dahl AA, Holsten F. Screening for postnatal depression. Validation of the Norwegian version of the Edinburgh Postnatal Depression Scale, and assessment of risk factors for postnatal depression. J Affective Disord. 2003;76:151-6.
10. Brodey BB, Goodman SH, Baldasaro RE, Brooks-DeWeese A, Wilson ME, Brodey ISB, Doyle NM. 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.

11. Chibanda D, Mangezi W, Tshimanga M, Woelk G, Rusakaniko P, Stranix-Chibanda L, et al. Validation of the Edinburgh Postnatal Depression Scale among women in a high HIV prevalence area in urban Zimbabwe. Arch Womens Ment Health. 2010;13:201-6.
12. Christl B, Reilly N, Smith M, Sims D, Chavasse F, Austin MP. The mental health of mothers of unsettled infants: is there value in routine psychosocial assessment in this context? Arch Womens Ment Health. 2013;16:391-9.
13. Crotty F, Sheehan J. Prevalence and detection of postnatal depression in an Irish community sample. Irish Journal of Psychological Medicine. 2004;21:117.
14. Gausia K, Fisher C, Algin S, Oosthuizen J. Validation of the Bangla version of the Edinburgh Postnatal Depression Scale for a Bangladeshi sample. J Reprod Infant Psychol. 2007;25:308-15.
15. Gorman LL, O'Hara MW, Figueiredo B, Hayes S, Jacquemain F, Kammerer MH, et al. Adaptation of the structured clinical interview for DSM-IV disorders for assessing depression in women during pregnancy and post-partum across countries and cultures. Br J Psychiatry Suppl. 2004;46:s17.
16. Li L, Liu F, Zhang H, Wang L, Chen X. Chinese version of the Postpartum Depression Screening Scale: translation and validation. Nurs Res. 2011;60:231-9.
17. Mahmud WM, Awang A, Mohamed MN. Revalidation of the Malay Version of the Edinburgh Postnatal Depression Scale (EPDS) Among Malay Postpartum Women Attending the Bakar Bata Health Center in Alor Setar, Kedah, North West Of Peninsular Malaysia. Malays J Med Sci. 2003;10:71-5.
18. Matthey S, Barnett B, Kavanagh DJ, Howie P. Validation of the Edinburgh Postnatal Depression Scale for men, and comparison of item endorsement with their partners. Journal of Affective Disorders. 2001;64:175.
19. Moses-Kolko EL, Price JC, Wisner KL, Hanusa BH, Meltzer CC, Berga SL, et al. Postpartum and depression status are associated with lower [11C]raclopride BPND in Reproductive-Age Women. Neuropsychopharmacol. 2012;37:1422-32.
20. O'Brien LM, Heycock EG, Hanna M, Jones PW, Cox JL. Postnatal depression and faltering growth: A community study Pediatrics. 2004;113:1242.

21. Pedersen C, Leserman J, Garcia N, Stansbury M, Meltzer-Brody S, Johnson J. Late pregnancy thyroid-binding globulin predicts perinatal depression. Psychoneuroendocrinology. 2016;65:84-93.
22. Pinheiro RT, Coelho FM, Silva RA, Pinheiro KA, Oses JP, Quevedo Lde A, et al. Association of a serotonin transporter gene polymorphism (5-HTTLPR) and stressful life events with postpartum depressive symptoms: a population-based study. J Psychosom Obstet Gynecol. 2013;34:29-3
23. Priest SR, Henderson J, Evans SF, Hagan R. Stress debriefing after childbirth: a randomised controlled trial. Medical Journal of Australia. 2003;178:542.
24. Stuebe AM, Grewen K, MeltzerBrody S. Association between maternal mood and oxytocin response to breastfeeding. J Womens Health. 2013;22:352.

eTable4. Frequencies of EPDS scores for cases and non-cases of major depression in the full IPDMA dataset

EPDS Score	N non-cases	N cases
0	1243	9
1	961	12
2	1080	15
3	995	12
4	992	21
5	901	25
6	863	45
7	735	45
8	714	55
9	674	53
10	537	54
11	418	113
12	363	104
13	292	114
14	211	117
15	186	134
16	139	119
17	93	100
18	85	97
19	50	85
20	29	59
21	30	64
22	10	50
23	8	34
24	7	32
25	8	17
26	2	14
27	3	9
28	1	8
29	0	8
30	0	1
Total	**11630**	**1625**

eTable5. Bias in accuracy estimates of sample–specific optimal cutoffs compared to the accuracy estimates from population optimal cutoff, stratified by the magnitude of optimal cutoff

Sample–specific optimal cutoff	Sample size = 100			Sample size = 200			Sample size = 500			Sample size = 1,000		
	Mean Difference (95% CI)			Mean Difference (95% CI)			Mean Difference (95% CI)			Mean Difference (95% CI)		
	N	Sens	Spec	N	Sens	Spec	N	Sens	Spec	N	Sens	Spec
≥ 5 to ≥ 8	137	16.0 (14.8 to 17.2)	-19.6 (-20.8 to -18.3)	84	13.9 (13.1 to 14.8)	-16.5 (-17.6 to -15.3)	24	10.8 (9.2 to 12.4)	-16.3 (-17.8 to -14.8)	6	10.6 (9.0 to 12.1)	-13.8 (-15.3 to -12.3)
≥ 9 or ≥ 10	225	11.0 (10.0 to 12.1)	-5.2 (-5.8 to -4.6)	289	7.7 (7.0 to 8.5)	-5.1 (-5.6 to -4.6)	282	5.0 (4.5 to 5.6)	-5.2 (-5.6 to -4.9)	222	4.8 (4.3 to 5.2)	-5.2 (-5.5 to -4.9)
≥ 11	303	6.3 (5.2 to 7.3)	0.8 (0.4 to 1.3)	347	3.9 (3.1 to 4.7)	0.4 (0.1 to 0.7)	530	1.1 (0.7 to 1.6)	0.2 (0.0 to 0.3)	705	0.8 (0.5 to 1.0)	0.0 (-0.1 to 0.1)
≥ 12 or ≥ 13	244	2.1 (0.8 to 3.4)	5.3 (4.8 to 5.8)	239	-1.1 (-2.2 to 0.1)	5.0 (4.7 to 5.4)	161	-3.0 (-3.8 to -2.1)	4.4 (4.1 to 4.8)	67	-4.1 (-4.9 to -3.2)	3.7 (3.3 to 4.1)
≥ 14 to ≥ 17	91	-6.3 (-8.9 to -3.7)	10.7 (10.2 to 11.2)	41	-8.1 (-11.0 to -5.2)	9.8 (9.1 to 10.4)	3	-9.4 (-10.1 to -8.8)	9.9 (4.1 to 15.8)	NA	NA	NA

N = Number of samples; Sens = Sensitivity, Spec = Specificity, NA = Not available

Sample-specific optimal cutoff refers to the cutoff with maximum Youden's J (sensitivity + specificity - 1) in the simulated sample

Population optimal cutoff refers to the EPDS cutoff of ≥ 11 that maximized Youden's J (sensitivity + specificity - 1) in the full IPDMA dataset