Journal of Clinical Epidemiology 137 (2021) 137–147

ORIGINAL ARTICLE

Data-driven methods distort optimal cutoffs and accuracy estimates of depression screening tools: a simulation study using individual participant data

Parash Mani Bhandari, MSc[a,b], Brooke Levis, PhD[a,b,c], Dipika Neupane, MSc[a,b], Scott B. Patten, PhD[d], Ian Shrier, MD[a,b,e], Brett D. Thombs, PhD[a,b,f,g,h,i,j,*], Andrea Benedetti, PhD[a,b,f,*], the Depression Screening Data (DEPRESSD) EPDS Group[#]

[a] *Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada*
[b] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada*
[c] *Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, UK*
[d] *Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada*
[e] *Department of Family Medicine, McGill University, Montréal, Québec, Canada*
[f] *Department of Medicine, McGill University, Montréal, Québec, Canada*
[g] *Department of Psychiatry, McGill University, Montréal, Québec, Canada*
[h] *Department of Psychology, McGill University, Montréal, Québec, Canada*
[i] *Department of Educational and Counselling Psychology, McGill University, Montréal, Québec, Canada*
[j] *Biomedical Ethics Unit, McGill University, Montréal, Québec, Canada*

Accepted 29 March 2021; Available online 8 April 2021

Abstract

Objective: To evaluate, across multiple sample sizes, the degree that data-driven methods result in (1) optimal cutoffs different from population optimal cutoff and (2) bias in accuracy estimates.

Study design and setting: A total of 1,000 samples of sample size 100, 200, 500 and 1,000 each were randomly drawn to simulate studies of different sample sizes from a database (n = 13,255) synthesized to assess Edinburgh Postnatal Depression Scale (EPDS) screening accuracy. Optimal cutoffs were selected by maximizing Youden's J (sensitivity+specificity–1). Optimal cutoffs and accuracy estimates in simulated samples were compared to population values.

Results: Optimal cutoffs in simulated samples ranged from $\geq$ 5 to $\geq$ 17 for n = 100, $\geq$ 6 to $\geq$ 16 for n = 200, $\geq$ 6 to $\geq$ 14 for n = 500, and $\geq$ 8 to $\geq$ 13 for n = 1,000. Percentage of simulated samples identifying the population optimal cutoff ($\geq$ 11) was 30% for n = 100, 35% for n = 200, 53% for n = 500, and 71% for n = 1,000. Mean overestimation of sensitivity and underestimation of specificity were 6.5 percentage point (pp) and -1.3 pp for n = 100, 4.2 pp and -1.1 pp for n = 200, 1.8 pp and -1.0 pp for n = 500, and 1.4 pp and -1.0 pp for n = 1,000.

Conclusions: Small accuracy studies may identify inaccurate optimal cutoff and overstate accuracy estimates with data-driven methods.

Keywords: Optimal cutoff; Accuracy estimates; Bias; Cherry-picking; Data-driven methods; Depression

Conflict of Interest: All authors have completed the ICMJE uniform disclosure form and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years with the following exceptions: Dr. 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.

[#] DEPRESSD EPDS Collaboration listed at end of article.

[*] Corresponding author. Brett D. Thombs: Tel. (514) 340-8222 ext. 25112; Andrea Benedetti: Tel. (514) 934-1934 ext. 32161.
E-mail addresses: brett.thombs@mcgill.ca (B.D. Thombs), andrea.benedetti@mcgill.ca (A. Benedetti).

https://doi.org/10.1016/j.jclinepi.2021.03.031
0895-4356/

What is new?

Key findings

- Optimal EPDS cutoffs identified in samples of different sizes varied widely, ranging from ≥ 5 to ≥ 17 for studies with $n = 100$ and ≥ 8 to ≥13 for $n = 1{,}000$.
- Mean overestimation of sensitivity and underestimation of specificity, respectively, were 6.5 percentage point (pp) and -1.3 pp for $n = 100$ and 1.4 pp and -1.0 pp for $n = 1{,}000$.

What this adds to what is known?

- This is the first study to use real patient data to estimate the degree that data-driven methods result in selection of inaccurate optimal cutoffs and bias in accuracy estimates.

What is the implication?

- Optimal cutoffs identified in primary accuracy studies are often incorrect and accuracy estimates are often overstated.
- Researchers should avoid making recommendations about cutoffs to use in practice and accuracy when reporting results from small single studies.
- Clinicians should select cutoffs for specific populations that are generated from well-conducted meta-analyses or identified consistently across multiple large primary studies.

1. Introduction

Depression screening tools are commonly used to identify patients with unrecognized and untreated depression [1,2]. Evidence from studies on the accuracy of depression screening tools is used to select an "optimal" cutoff for use in practice and to estimate accuracy for distinguishing between positive and negative results using that cutoff. Many studies on depression screening tools, however, use data-driven approaches, by which investigators use the same dataset to select an optimal cutoff and estimate screening accuracy at that cutoff. Cutoffs selected in this way may deviate substantially from a true population optimal cutoff that would be selected if a sufficiently large or population database were available. Additionally, accuracy estimates generated in these studies may be optimistic compared to what would occur in clinical practice.

The Edinburgh Postnatal Depression Scale (EPDS) is the most commonly used tool to screen for depression during pregnancy and postpartum [3,4]. Diagnosis of depression in pregnancy and postpartum is particularly challenging, since some symptoms overlap with normal experiences during this period, such as loss of appetite, poor sleep and fatigue [5-7]. Some health care practitioners may not fully understand the information provided by EPDS; different resources suggest that cutoffs of ≥ 10 and ≥ 13 can be used to identify women with "possible" or "probable" depression [8,9], but only approximately 35% and 60% of women who score above cutoffs of ≥ 10 and ≥ 13, respectively, will experience a major depressive episode, assuming a prevalence of 10% [10,11].

Misunderstanding about how to interpret results from depression screening tools is compounded by results from primary studies that suggest that different cutoffs identified as "optimal" in their studies should be used in specific populations. These results are often generated using data-driven analytical approaches in small samples. In many of these studies, the abstract, which may be the only part of the article that is read [12,13], only reports accuracy results from a single data-driven optimal cutoff rather than from standard cutoffs. Data-driven methods that are sometimes used for optimal cutoff selection include selecting the cutoff that maximizes Youden's J (sensitivity + specificity–1), minimizes Euclidean distance (distance to the corner of the receiver operator characteristic curve) or maximizes diagnostic odds ratio [14]. Youden's J is the method most frequently used in diagnostic accuracy studies for depression screening tools. Illustrating this, we reviewed recently published primary studies of EPDS accuracy (N participants, range = 118-807; mean = 320) and found that only 1 of 14 studies (7%) reported accuracy results for more than one cutoff in the abstract. The remaining 13 (93%) only reported results from the single best-performing cutoff, which was based on maximizing Youden's J in 11 of the 13 (85%) studies. Cutoffs identified as optimal in the 14 studies ranged from ≥ 8 to ≥ 13. In many of the studies, when data-driven optimal cutoffs diverged from more standard cutoffs, authors suggested that this represented a unique optimal cutoff that should be used in the study's specific target population group. No studies attributed a divergent optimal cutoff to a small sample size or to data-driven cutoff selection methods (Appendix-eMethods1).

We know of only four studies that have investigated the degree to which data-driven selection of cutoff may influence diagnostic accuracy estimates [15-18]. These studies each reported that data-driven cutoff selection produces overly optimistic estimates, particularly in small samples. However, these studies used simulated datasets based on hypothetical test score distributions rather than real participant data. Thus, how widely data-driven optimal cutoffs diverge from population-based optimal cutoffs and how biased estimates of diagnostic accuracy may be based on actual participant data is not known for any depression screening test, including the EPDS.

The objectives of the present study were to illustrate for users of evidence on depression screening tool accuracy, such as the EPDS, across different study sample sizes, 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 a

"population" dataset and (2) generates biased accuracy results compared to results from the population optimal cutoff.

2. Methods

We used a database originally synthesized for an individual participant data meta-analysis (IPDMA) on the accuracy of the EPDS for depression screening to form a study population from which we simulated studies of different sample sizes [10]. A protocol for the present study was uploaded to the Open Science Framework repository prior to initiating the study (https://osf.io/qnvzp/).

Details on methodology for the original IPDMA used in this study are published elsewhere [10], and are provided in Appendix-eMethods2.

2.1. Simulation of study samples and statistical analyses

Unlike many other depression screening tools, 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 [6]. However, many studies report using different cutoffs between ≥ 10 and ≥ 13 to identify major depression [19,20], with ≥ 13 being the most common [20]. A recent IPDMA using an updated and slightly larger version of the dataset used in the present study found that a cutoff of ≥ 11 maximized Youden's J overall and for subgroups.

For the present study, we used our IPDMA dataset to represent a hypothetical "population" of women, and 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 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 IPDMA, which used sampling weights and study-level clustering and a slightly larger sample. We verified that a cutoff of ≥ 11 maximized Youden's J for the unweighted population.

From the population IPDMA dataset, we sampled with replacement to generate 1,000 random samples of sample size 100, 200, 500, 1,000 each. 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. For each sample size, across the 1,000 samples, we (1) graphically illustrated the variability in sample-specific optimal cutoffs and the variability in accuracy of the sample-specific optimal cutoffs; (2) estimated the mean difference (bias) 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 (bias) 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 2×2 table counts. In additional analyses, we stratified results by the optimal cutoff value identified in each sample.

2.2. 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 [19,20]. We subsequently determined that a population optimal cutoff of ≥ 11 maximizes Youden's J in our IPDMA "population", which was established in the main IPDMA database and confirmed in the present study. Thus, we used a cutoff of ≥ 11 only and not ≥ 13, since the purpose was to determine how data-driven results would diverge from similar analyses done with population data.

3. Results

The original IPDMA database included 49 primary studies with 13,255 participants (1,625 major depression cases, 12.3%), which constituted the "population" for the present study. Characteristics of the primary studies included in the IPDMA database are provided in Appendix-eTable1. The sample sizes of the primary studies ranged from 40 to 2,634 (mean = 271, median = 190). The mean number of cases of major depression was 34 (median = 25), and 20 studies included < 20 cases of major depression. Frequencies of EPDS scores for cases and non-cases in the IPDMA database are shown in Appendix-eTable2. 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, 80.7) and 83.4% (95% CI: 82.7, 84.0).

3.1. Variability of sample-specific optimal cutoffs in simulated samples

Figure 1 shows the variability of sample-specific optimal cutoffs for each sample size. Optimal cutoffs in indi-

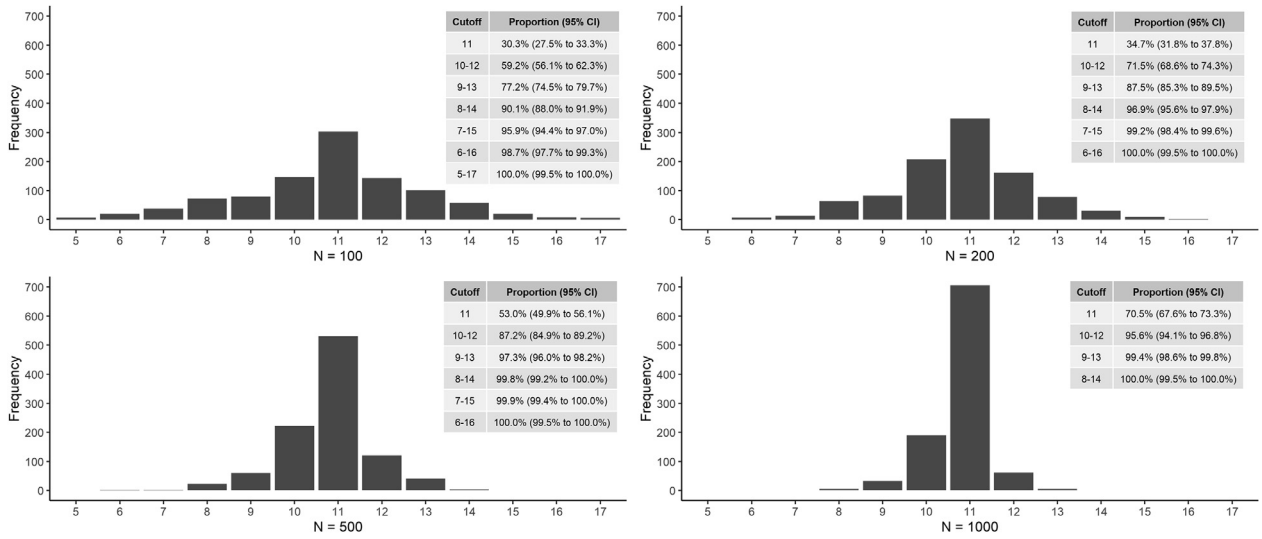

Fig. 1. Variability in optimal cutoffs in 1,000 simulated samples of sample size 100, 200, 500 and 1,000.
Optimal cutoff was defined as the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the study sample. 26 out of the 4,000 simulated samples had a tie in maximum Youden's J, and the higher cutoff was selected as the optimal cutoff.

Table 1. Bias in accuracy estimates (in percentage point) in 1,000 simulated samples of sample 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, 7.2)	-1.3 (-1.9, -0.7)	4.2 (3.6, 4.7)	-1.1 (-1.6, -0.7)	1.8 (1.4, 2.1)	-1.0 (-1.3, -0.7)	1.4 (1.1, 1.6)	-1.0 (-1.2, -0.8)
Sample cutoff ≥ 11 values – Population values with cutoff ≥ 11	0.0 (-0.7, 0.8)	0.1 (-0.2, 0.3)	0.2 (-0.3, 0.8)	0.1 (-0.1, 0.3)	-0.2 (-0.6, 0.1)	0.0 (-0.1, 0.1)	0.1 (-0.1, 0.3)	-0.1 (-0.1, 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.

vidual 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 $n = 1{,}000$. There was a tie in maximum Youden's J between multiple cutoffs in 26 of the 4,000 samples. The percentage of samples that identified the true population optimal cutoff of ≥ 11 was 30.3% (95% CI: 27.5, 33.3) for $n = 100$, 34.7% (95% CI: 31.8, 37.8) for $n = 200$, 53.0% (95% CI: 49.9, 56.1) for $n = 500$, and 70.5% (95% CI: 67.6, 73.3) for $n = 1{,}000$.

3.2. Bias from data-driven cutoff selection in simulated samples

As shown in 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 percentage point (pp) (95% CI: 5.8, 7.2) for $n = 100$ [i.e. mean sensitivity of optimal cutoffs in 1,000 simulated samples of $n = 100$ (85.2%) – true population sensitivity (78.7%) = 6.5 pp], 4.2pp (95% CI: 3.6, 4.7) for $n = 200$, 1.8 pp (95% CI: 1.4, 2.1) for $n = 500$ and 1.4 pp (95% CI: 1.1, 1.6) for $n = 1{,}000$. Specificity was underestimated by 1.3 pp (95% CI: -1.9, -0.7) for $n = 100$, 1.1 pp (95% CI: -1.6, -0.7) for $n = 200$, 1.0 pp (95% CI: -1.3, -0.7) for $n = 500$ and 1.0 pp (95% CI: -1.2, -0.8) for $n = 1{,}000$. Figure 2 presents quartiles of the accuracy estimates for simulated samples.

Figure 3 and Appendix-eTable3 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 $n = 100$, in samples with sample-specific optimal cutoff ≥ 5 to ≥ 8, sensitivity was overestimated by 16.0 pp (95% CI: 14.8, 17.2), and specificity was underestimated by 19.6 pp (95% CI: -20.8, -

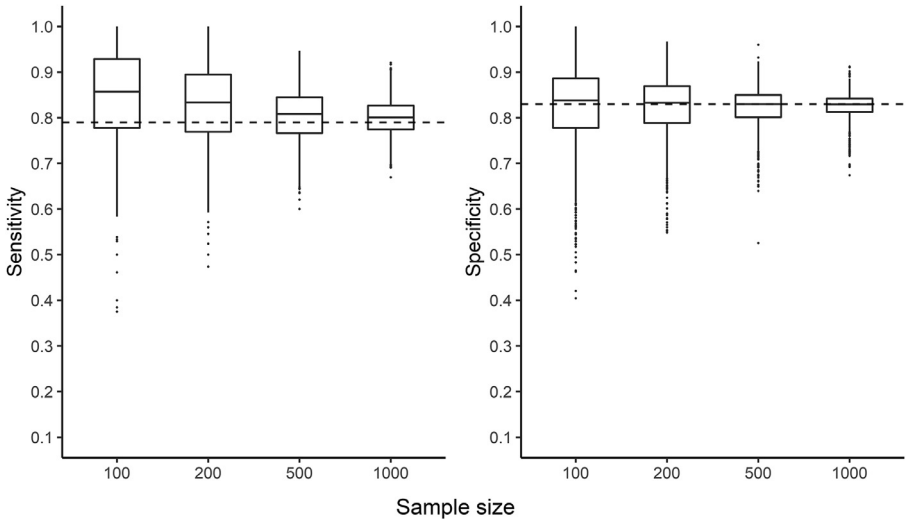

Fig. 2. Boxplots of accuracy estimates of the optimal cutoff in 1,000 simulated samples of sample size 100, 200, 500 and 1,000 compared to the accuracy estimates of cutoff ≥ 11 in the population.
Optimal cutoff refers to the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the study sample. Dotted horizontal line represents the accuracy of cutoff ≥ 11 in the population (full IPDMA dataset). Boxplots present quartiles (first quartile, median and third quartile) of accuracy estimates for the simulated samples.

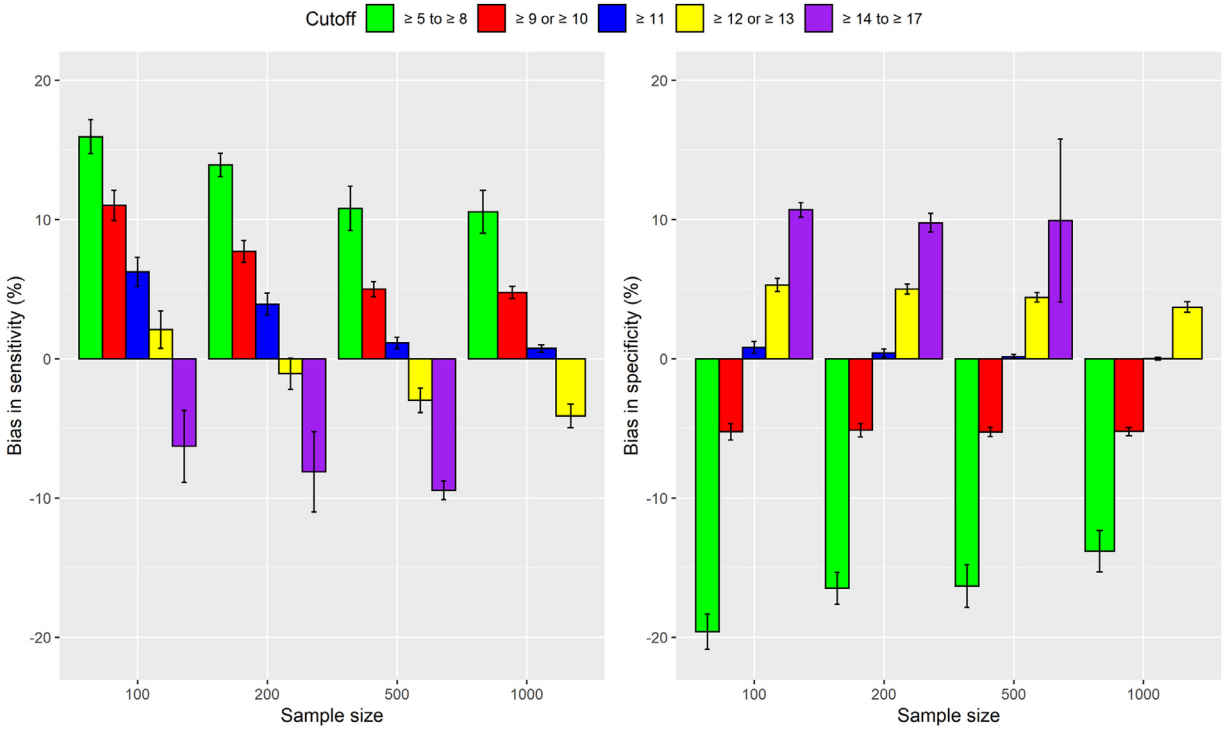

Fig. 3. Bias in accuracy estimates in simulated samples of sample size 100, 200, 500 and 1,000 stratified by sample optimal cutoffs.
Optimal cutoff refers to the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the study sample. The error bars represent 95% confidence intervals of the bias in accuracy estimates.

18.3). For samples with sample-specific optimal cutoffs of ≥ 14 to ≥ 17, sensitivity was underestimated by 6.3 pp (95% CI: -8.9, -3.7), and specificity was overestimated by 10.7 pp (95% CI: 10.2, 11.2).

As shown in Figure 4, 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. See also Table 1.

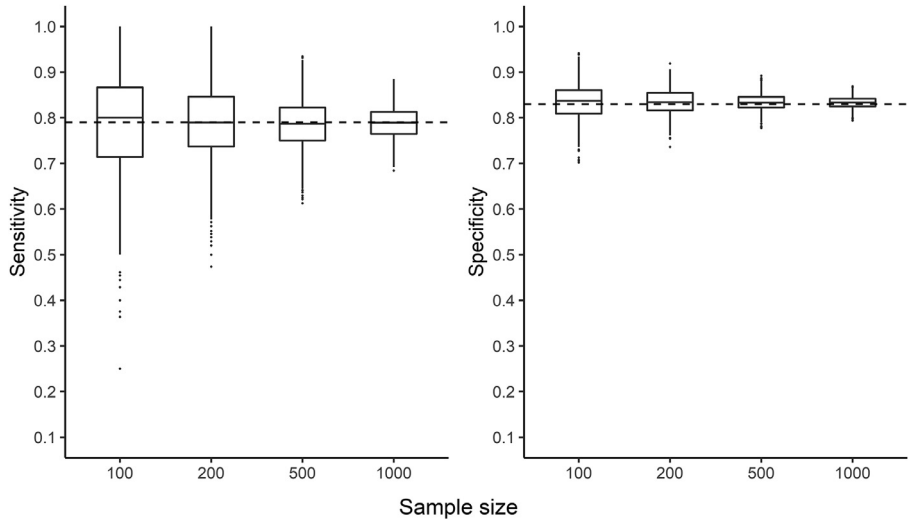

Fig. 4. Boxplots of accuracy estimates of the cutoff ≥ 11 in 1,000 simulated samples of sample size 100, 200, 500 and 1,000 compared to the accuracy estimates of cutoff ≥ 11 in the population.
Optimal cutoff refers to the cutoff that maximized Youden's J (sensitivity + specificity – 1) in the sample. Dotted horizontal line represents the accuracy of cutoff ≥ 11 in the population (full IPDMA dataset). Boxplots present quartiles (first quartile, median and third quartile) of accuracy estimates for the simulated samples.

4. Discussion

There were two main findings of this study. First, with very small sample sizes ($n = 100$), study-specific optimal cutoffs ranged from ≥ 5 to ≥ 17 (compared to the actual population optimal cutoff of ≥ 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 pp, whereas mean underestimation of specificity was 1.3 pp. For larger samples ($n = 1{,}000$), sensitivity was overestimated, on average, by 1.4 pp and specificity underestimated by 1.0 pp. The degree and direction of bias 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 16.0pp; studies that identified high optimal cutoffs (≥ 14 to ≥ 17), on the other hand, underestimated sensitivity by 6.3 pp.

The degree of variability 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 database, 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 1,000 and only one (2.0%) had sample size > 1,000. 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 1,000 and 7 (7.9%) had sample size of > 1,000 [21]. Based on our findings, overall, many studies of depression screening tool accuracy likely overestimate sensitivity with only minor losses in specificity. A larger bias in sensitivity estimates as compared to bias in specificity estimates is intuitive, as most studies have much fewer participants with major depression (among whom sensitivity is estimated) than without (among whom specificity is estimated). Thus, optimal cutoff selection in some samples can result 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.

Surveys have shown that clinicians have difficulty understanding medical statistics, including conditional probabilities such as sensitivity, specificity, positive predictive value and negative predictive value [22-24]. Thus, clinicians may misinterpret EPDS cutoffs with inflated sensitivity estimates from data-driven procedures as being virtually diagnostic, and adopt such cutoffs for use in clinical practice, even when the actual positive predictive value may be much smaller [25].

Clinicians who use the EPDS in their practice should be wary of EPDS cutoff recommendations based on small individual studies that used data-driven methods to identify the optimal cutoff. Such cutoffs are likely to not truly be optimal for the population of interest, and accuracy estimates are likely to be overly optimistic compared to what would be obtained in actual clinical practice. Instead, clin-

icians should select EPDS cutoff thresholds from large, well-conducted meta-analyses or validated across multiple studies. In addition, clinicians may also opt to prioritize either sensitivity or specificity in different clinical settings, and select higher or lower thresholds, depending on their health and financial priorities.

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 [26]. 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 three reported *a priori* sample size calculations, and none specifically considered the issue of identifying an optimal cutoff and estimating accuracy in the same participant sample [21]. Authors of primary studies on depression screening tool accuracy could potentially use statistical methods to estimate confidence intervals for uncertainty around the optimal cutoff [27,28]. They could also employ internal validation methods such as cross-sampling, sample-splitting and bootstrapping to statistically adjust for the bias in accuracy estimates from data-driven optimal cutoff selection [29]. 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 [30,31]. 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 [32].

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. Since the same method was used in this study to identify the population optimal cutoff and the optimal cutoff in each simulated study sample, the reason why study sample optimal cutoffs diverge, particularly in very small samples is due to the sampling distribution of the mean in relatively small samples. Hence, authors of individual studies should avoid recommending specific cutoffs for specific populations unless the studies use very large samples, or the findings are replicated consistently across multiple studies.

Strengths of this study include the use of real-participant data instead of simulated data and hypothetical distributional assumptions and the large population from which we were able to draw samples. There are also limitations to consider. Our results on the bias in accuracy 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 study sample. Another possible limitation is that we only used Youden's J to select optimal cutoffs. It is the most commonly used method, by far, in depression screening accuracy studies and performs similarly to other indices, such as the Euclidean distance [14]. It is possible, however, that results might slightly differ for an alternative method for selecting optimal cutoffs.

5. Conclusions

We found that data-driven cutoff selection methods often result in optimal cutoffs that differ from the population optimal cutoff and in accuracy estimates that are overly optimistic. 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; and should report results completely. Clinicians should be aware that cutoffs and accuracy from single studies may not reflect what will occur in practice and should select a cutoff from well-conducted meta-analyses or identified consistently across multiple studies.

The Depression Screening Data (DEPRESSD) EPDS Group

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

ing, McGill University, Montréal, Québec, Canada; Pim Cuijpers, Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit, Amsterdam, the Netherlands; Simon Gilbody, Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK; John P.A. Ioannidis, Department of Medicine, Department of Health Research and Policy, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, California, USA; Lorie A. Kloda, Library, Concordia University, Montréal, Québec, Canada; Roy C. Ziegelstein, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; Liane Comeau, International Union for Health Promotion and Health Education, École de santé publique de l'Université de Montréal, Montréal, Québec, Canada; Nicholas D. Mitchell, Department of Psychiatry, University of Alberta, Edmonton, Alberta, Canada; Marcello Tonelli, Department of Medicine, University of Calgary, Calgary, Alberta, Canada; Simone N. Vigod, Women's College Hospital and Research Institute, University of Toronto, Toronto, Ontario, Canada; Franca Aceti, Department of Neurology and Psychiatry, Sapienza University of Rome, Rome, Italy; Rubén Alvarado, School of Public Health, Faculty of Medicine, Universidad de Chile, Santiago, Chile; Cosme Alvarado-Esquivel, Laboratorio de Investigación Biomédica, Facultad de Medicina y Nutrición, Avenida Universidad, Dgo, Mexico; Muideen O. Bakare, Child and Adolescent Unit, Federal Neuropsychiatric Hospital, Enugu, Nigeria; Jacqueline Barnes, Department of Psychological Sciences, Birkbeck, University of London, UK; Amar D. Bavle, Department of Psychiatry, Rajarajeswari Medical College and Hospital, Bengaluru, Karnataka, India; Cheryl Tatano Beck, University of Connecticut School of Nursing, Mansfield, Connecticut, USA; Carola Bindt, Department of Child and Adolescent Psychiatry, University Medical Center Hamburg-Eppendorf, Germany; Philip M. Boyce, Discipline of Psychiatry, Westmead Clinical School, Sydney Medical School, University of Sydney, Sydney, Australia; Adomas Bunevicius, Neuroscience Institute, Lithuanian University of Health Sciences, Kaunas, Lithuania; Tiago Castro e Couto, Federal University of Uberlândia, Brazil; Linda H. Chaudron, University of Rochester School of Medicine and Dentistry, Rochester, New York, USA; Humberto Correa, Medicine Faculty - Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brazil; Felipe Pinheiro de Figueiredo, Department of Neurosciences and Behavior, Ribeirão Preto Medical School, Brazil; Valsamma Eapen, University of New South Wales and Ingham Institute South West Sydney LHD, Australia; Nicolas Favez, Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland; Ethel Felice, Department of Psychiatry, Mount Carmel Hospital, Attard, Malta; Michelle Fernandes, Faculty of Medicine, Department of Paediatrics, University of Southampton, Southampton and Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK; Barbara Figueiredo, School of Psychology, University of Minho, Portugal; Jane R. W. Fisher, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; Lluïsa Garcia-Esteve, Perinatal Mental Health Unit CLINIC-BCN, Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain; Lisa Giardinelli, Psychiatry Unit, Department of Health Sciences, University of Florence, Firenze, Italy; Nadine Helle, Department of Child and Adolescent Psychiatry, University Medical Center Hamburg-Eppendorf, Germany; Louise M. Howard, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; Dina Sami Khalifa, Ahfad University for Women, Omdurman, Sudan; Jane Kohlhoff, University of New South Wales, Kensington, Australia; Zoltán Kozinszky, Department of Obstetrics and Gynaecology, Danderyd Hospital, Stockholm, Sweden; Laima Kusminskas, Private Practice, Hamburg, Germany; Lorenzo Lelli, Psychiatry Unit, Department of Health Sciences, University of Florence, Firenze, Italy; Angeliki A. Leonardou, First Department of Psychiatry, Women's Mental Health Clinic, Athens University Medical School, Athens, Greece; Michael Maes, Department of Psychiatry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand; Valentina Meuti, Department of Neurology and Psychiatry, Sapienza University of Rome, Rome, Italy; Sandra Nakić Radoš, Department of Psychology, Catholic University of Croatia, Zagreb, Croatia; Purificación Navarro García, Perinatal Mental Health Unit CLINIC-BCN. Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain; Daisuke Nishi, Department of Mental Health, Graduate School of Medicine, The University of Tokyo, Japan; Daniel Okitundu Luwa E-Andjafono, Unité de Neuropsychologie, Département de Neurologie, Centre Neuro-psycho-pathologique, Faculté de Médecine, Université de Kinshasa, République Démocratique du Congo; Susan J. Pawlby, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; Chantal Quispel, Department of Obstetrics and Gynaecology, Albert Schweitzer Ziekenhuis, Dordrecht, the Netherlands; Emma Robertson-Blackmore, Halifax Health, Graduate Medical Education, Daytona Beach, FL. USA; Tamsen J. Rochat, MRC/Developmental Pathways to Health Research Unit, School of Clinical Medicine, University of Witwatersrand, South Africa; Heather J. Rowe, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; Deborah J. Sharp, Centre for Academic Primary Care, Bristol Medical School, University of Bristol, UK; Bonnie W. M. Siu, Department of Psychiatry, Castle Peak Hospital, Hong Kong SAR, China; Alkistis Skalkidou, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden; Alan Stein, University of Oxford, Oxford, UK; Robert C. Stewart, Department of Mental Health, College of Medicine, University of Malawi, Malawi; Kuan-Pin Su, College of Medicine, China Medical

University, Taichung, Taiwan; Inger Sundström-Poromaa, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden; Meri Tadinac, Department of Psychology, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia; S. Darius Tandon, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; Iva Tendais, School of Psychology, University of Minho, Portugal; Pavaani Thiagayson, Institute of Mental Health, Singapore; Annamária Töreki, Department of Emergency, University of Szeged, Hungary; Anna Torres-Giménez, Perinatal Mental Health Unit CLINIC-BCN. Institut Clínic de Neurociències, Hospital Clínic, Barcelona, Spain; Thach D. Tran, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; Kylee Trevillion, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; Katherine Turner, Epilepsy Center-Child Neuropsychiatry Unit, ASST Santi Paolo Carlo, San Paolo Hospital, Milan, Italy; Johann M. Vega-Dienstmaier, Facultad de Medicina Alberto Hurtado, Universidad Peruana Cayetano Heredia. Lima, Perú; Karen Wynter, School of Nursing & Midwifery, Deakin University, Melbourne, Australia; and Kimberly A. Yonkers, Department of Psychiatry, Yale School of Medicine, New Haven, Connecticut, USA.

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 Nakić 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). Dr. Thombs was supported by a Tier 1 Canada Research Chair. Dr. Benedetti was supported by FRQ-S Researcher Salary Awards. No other authors reported funding for primary studies or for their work on the present study.

Author Contributions

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.

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.

Data Sharing

Requests to access data should be made to the corresponding authors.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.03.031.

References

[1] Siu AL. U S Preventive Services Task Force. Screening for depression in children and adolescents: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2016;164:360–6.
[2] Siu AL, Bibbins-Domingo K, Grossman DC, Baumann LC, Davidson KW, Ebell M. Screening for depression in adults: US Preventive Services Task Force recommendation statement. JAMA 2016;315:380–7.
[3] Hewitt CE, Gilbody SM, Mann R, Brealey S. 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.
[4] Howard LM, Molyneaux E, Dennis CL, Rochat T, Stein A, Milgrom J. Non-psychotic mental disorders in the perinatal period. Lancet 2014;384:1775–88.
[5] Klein MH, Essex MJ. Pregnant or depressed? The effect of overlap between symptoms of depression and somatic complaints of pregnancy on rates of major depression in the second trimester. Depression 1994;2:308–14.
[6] 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.
[7] Striegel-Moore R, Goldman SL, Garvin V, Rodin J. A prospective study of somatic and emotional symptoms of pregnancy. Psychol Women Q 1996;20:393–408.

[8] Stevenson K, Alameddine R, Rukbi G, Chahrouri M, Usta J, Saab B. High rates of maternal depression amongst Syrian refugees in Lebanon - a pilot study. Sci Rep 2019;9 11849-11019.
[9] Gibson J, McKenzie-McHarg K, Shakespeare J, Price J, Gray R. A systematic review of studies validating the Edinburgh Postnatal Depression Scale in antepartum and postpartum women. Acta Psychiatr Scand 2009;119:350–64.
[10] Levis B, Negeri Z, Sun Y, Benedetti A, Thombs BD. DEPRESsion Screening Data (DEPRESSD) EPDS Group. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. BMJ 2020;371:m4022.
[11] Vesga-López O, Blanco C, Keyes K, Olfson M, Grant BF, Hasin DS. Psychiatric disorders in pregnant and postpartum women in the United States. Arch Gen Psychiatry 2008;65:805–15.
[12] Pitkin RM, Branagan MA. Can the accuracy of abstracts be improved by providing specific instructions? A randomized controlled trial. JAMA 1998;280:267–9.
[13] Beller EM, Glasziou PP, Altman DG, Hopewell S, Bastian H, Chalmers I. PRISMA for Abstracts: reporting systematic reviews in journal and conference abstracts. PLoS Med 2013;10:e1001419.
[14] Hajian-Tilaki K. The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. Stat Methods Med Res 2020;27:2374–83.
[15] Linnet K, Brandt E. Assessing diagnostic tests once an optimal cut-off point has been selected. Clin Chem 1986;32:1341–6.
[16] Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol 2006;59:798–801.
[17] 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.
[18] Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem 2008;54:729–37.
[19] O'Connor E, Rossom RC, Henninger M, Groom HC, Burda BU. 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.
[20] Hewitt C, Gilbody S, Brealey S, Paulden M, Palmer S, Mann R. Methods to identify postnatal depression in primary care: an integrated evidence synthesis and value of information analysis. Health Technol Assess 2009;13:1–230.
[21] 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.
[22] Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. Am J Med 1981;71:991–8.
[23] Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychol Sci Public Interest 2007;8:53–96.
[24] Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA 2007;298:1010–22.
[25] Aranake-Chrisinger A, Avidan MS. Don't crush the sensitive snout. Reply to: Sensitivity is not enough. Br J Anaesth 2017;119:1240–1.
[26] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.
[27] Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J 2005;47:458–72.
[28] Schisterman EF, Perkins N. Confidence intervals for the Youden Index and corresponding optimal cut-point. Commun Stat Simul Comput 2007;36:549–63.
[29] Smith GC, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. Am J Epidemiol 2014;180:318–24.
[30] Thombs BD, Benedetti A, Kloda LA, Levis B, Riehm KE, Azar M. 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-e002015.
[31] Thombs BD, Benedetti A, Kloda LA, Levis B, Nicolau I, Cuijpers P. 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.
[32] Levis B, Benedetti A, Levis AW, Ioannidis JP, Shrier I, Cuijpers P. 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.