Journal of Affective Disorders 205 (2016) 335–343

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Review article

Summary diagnostic validity of commonly used maternal major depression disorder case finding instruments in the United States: A meta-analysis

Arthur H. Owora *, Hélène Carabin, Jessica Reese, Tabitha Garwe

Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States

ARTICLE INFO

Article history:
Received 16 May 2016
Received in revised form
5 July 2016
Accepted 14 August 2016
Available online 16 August 2016

Keywords:
Diagnostic performance
Case-finding instrument
Major depression disorder
Bayesian meta-analysis
Misclassification error

ABSTRACT

Introduction: Major Depression Disorder (MDD) is common among mothers of young children. However, its detection remains low in primary-care and community-based settings in part due to the uncertainty regarding the validity of existing case-finding instruments. We conducted meta-analyses to estimate the diagnostic validity of commonly used maternal MDD case finding instruments in the United States.
Methods: We systematically searched three electronic bibliographic databases PubMed, PsycINFO, and EMBASE from 1994 to 2015 to identify relevant published literature. Study eligibility and quality were evaluated using the Standards for the Reporting of Diagnostic Accuracy studies and Quality Assessment of Diagnostic Accuracy Studies guidelines, respectively. Pooled sensitivity and specificity of case-finding instruments were generated using Bayesian hierarchical summary receiver operating models.
Results: Overall, 1130 articles were retrieved and 74 articles were selected for full-text review. Twelve articles examining six maternal MDD case-finding instruments met the eligibility criteria and were included in our meta-analyses. Pooled sensitivity and specificity estimates were highest for the BDI-II (91%; 95% Bayesian Credible Interval (BCI): 68%; 99% and 89%; 95% BCI: 62%; 98% respectively) and EPDS10 (74%; 95% BCI: 46%; 91% and 97%; 95% BCI: 84%; 99% respectively) during the antepartum and postpartum periods respectively.
Limitation: No meta-regression was conducted to examine the impact of study-level characteristics on the results.
Discussion: Diagnostic performance varied among instruments and between peripartum periods. These findings suggest the need for a judicious selection of maternal MDD case-finding instruments depending on the study population and target periods of assessment.

Contents

* Correspondence to: 745 Martina Lane, Edmond, OK 73034, United Sates.
E-mail addresses: hamieuga@gmail.com (A.H. Owora), Helene-Carabin@ouhsc.edu (H. Carabin), Jessica-Reese@ouhsc.edu (J. Reese), Tabitha-Garwe@ouhsc.edu (T. Garwe).

http://dx.doi.org/10.1016/j.jad.2016.08.014
0165-0327/

1. Background

Major depression disorder (MDD) case-finding instruments rely on subjective symptoms, patient experiences and perceptions that are typically validated in the absence of a 'gold standard'. The sensitivity and specificity estimates of these instruments are based on comparing their classification to that of reference standards, which themselves include classification error. Reference standard errors result in biased case-finding instrument diagnostic performance estimates. Among mothers, during the peripartum period which includes antepartum and postpartum periods, the potential for case-finding instrument misclassification error (especially false positives) is likely to be heightened by the presence of 'morning sickness', 'baby blues' and parenting stress symptoms that mimic those of MDD (Pereira et al., 2014). These issues contribute in part to the uncertainty regarding how valid existing maternal MDD case-finding instruments are in detecting true MDD. As a consequence of this uncertainty, various stakeholders in the United States (i.e. US Preventive Services Task Force, American Congress of Obstetricians and Gynecologists, American Academy of Pediatrics, American Academy of Family Physicians and the American College of Nurse Midwives) have recommended inconsistent maternal MDD screening/case-finding practices and policies (Gaynes et al., 2005; Agency for Healthcare Research and Quality AHRQ, 2014; Pignone et al., 2002; O'Connor et al., 2016).

In order to address the uncertainty around the diagnostic validity of maternal MDD case-finding instruments, meta-analyses can be used to generate summary measures of the sensitivity and specificity based on studies deemed to be valid and comparable while maximizing precision estimates. Unfortunately, previous diagnostic validity systematic reviews of maternal MDD case-finding instruments have not generated instrument-specific and/or peripartum period-specific pooled sensitivity and specificity estimates largely due to variability in not only the populations studied, but also in the diagnostic thresholds and reference standards used (Gaynes et al., 2005; Agency for Healthcare Research and Quality AHRQ, 2014; Pignone et al., 2002; O'Connor et al., 2016). Furthermore, because both existing maternal MDD case-finding instruments and reference standards aim to at least partly meet the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria for MDD diagnosis, there could be conditional dependence in the errors (i.e. false positives and negatives) generated by these tests when used on the same individuals. Combined, these issues preclude definitive conclusions regarding the diagnostic validity of maternal MDD case-finding instruments.

Meta-analysis techniques that account and adjust for the above issues exist; (Sadatsafavi et al., 2010; Chu et al., 2009; Walter et al., 1999; Dendukuri et al., 2012; Bernatsky et al., 2005; Dendukuri and Joseph, 2001) however, such methods have not yet been applied to maternal MDD diagnostic accuracy studies. The objective of this study was to conduct meta-analyses to estimate the diagnostic validity of commonly used maternal MDD case finding instruments in the US while accounting for 1) varying diagnostic thresholds, 2) use of multiple imperfect reference standards to validate the same case-finding instrument, 3) and the potential for conditional dependence of errors generated from case-finding instrument and reference standard results.

2. Methods and procedures

2.1. Data sources and searches

Three electronic databases PubMed, PsycINFO, and EMBASE were searched for studies published from January 1st, 1994 to December 31st, 2015. An experienced librarian guided all searches. Older literature was excluded due to the publication of the Diagnostic *and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) in 1994. Briefly, our search strategy included various terms for depression, diagnostic performance, and the names of existing published MDD case-finding instruments and reference standards. To identify additional studies, we reviewed the bibliographies of included articles and previous systematic reviews.

2.2. Phase I – screening of abstracts

Titles and abstracts of identified studies were reviewed by Arthur Owora (AO) for further consideration. AO reviewed all articles without abstracts in full. Two exclusion criteria were used in Phase I: (1) no assessment of MDD, and (2) absence of original data. All articles not meeting these exclusion criteria were reviewed in Phase II.

2.3. Phase II – review of full articles

Articles moved to Phase II were reviewed in full using the following eligibility criteria (eTable 1a): MDD measured among mothers of young children (0–5 years old) in the US and reporting of both case-finding and reference standard instrument results. Articles that included mothers from other countries or mothers with only older children (>5 years) were excluded. Included articles were moved to Phase III for a qualitative review and quantitative data extraction.

2.4. Phase III – qualitative assessment and quantitative data extraction

Articles eligible for Phase III were evaluated for their epidemiological quality by two investigators (JR and AO). The investigators

answered 11 signaling questions to rate four Quality Assessment of Diagnostic Accuracy Studies - second version (QUADAS-2) criteria domains namely: (Whiting et al., 2011) 1) patient selection - three questions; 2) index test (i.e. case-finding instruments) – four questions; 3) reference standard – three questions; 4) flow/timing of assessments – four questions. Two additional signaling questions (not covered by the QUADAS-2 tool) were added to assess the potential for confounding and effect modification. Here, confounding refers to the distortion of the relationship between case-finding instrument and reference standard results due to a third variable (e.g. age, race, and peripartum period of assessment) whereas effect modification refers to that relationship changing depending on the levels or categories of the third variable. Our study only examined the potential confounding and effect modification effect of peripartum periods. We were unable to assess the impact of other factors (e.g. age, race) due to their inconsistent assessment and reporting across studies selected in Phase III.

Each signaling question was answered by yes/no/unclear and used to classify the likelihood bias as being low/high/uncertain. Details of each domain's assessment criteria and overall study quality ratings are summarized in eTable 1b and 1c, respectively. A study with a low risk of bias classification for all four, three or two and one or none of QUADAS-2 domains was assigned a 'good', 'fair' and 'poor' overall study quality rating, respectively.

Two investigators (AO and JR) extracted data elements recommended by the Standards for Reporting Diagnostic Accuracy studies (STARD) guidelines (Bossuyt et al., 2003) from all articles assigned a 'fair' or 'good' overall study quality rating. The data elements include the description of the 1) study participants; 2) study designs; 3) case-finding instruments and reference standards; 4) data collection procedures; 5) statistical methods; 6) contingency tables of the case-finding instruments compared to the reference standards used as the 'Gold Standard' and reported as True positives (TP), False Positive (FP), True Negatives (TN) and False Negatives (FN); 7) how missing and indeterminate results were handled; and 8) study limitations and external validity. Study authors were contacted for additional information if needed.

2.5. Statistical analysis

Sensitivity, specificity, and 95% confidence intervals for each diagnostic threshold and period of assessment (i.e. antepartum and postpartum) for case-finding instruments included in Phase III were estimated using the Meta-analysis of Diagnostic Accuracy studies (MADA) package in *R* 3.1.2 (R Development Core, 2010) (R Development Core Team, 2015; Philipp Doebler, 2015). Antepartum depression was defined as an episode of MDD with onset occurring during pregnancy. The term 'episode' here refers to any two-week period during which depressive symptoms experienced by an individual meet DSM-IV MDD diagnostic criteria (Pereira et al., 2014). Postpartum depression was defined as an episode of MDD with the onset of symptoms occurring after childbirth (range: 1–14 months).

Systematic patterns in scatter plots of sensitivity and specificity estimates were examined to identify study population characteristics that influence instrument diagnostic performance. We examined four study participant characteristics namely the peripartum period of assessment; the trimester of pregnancy; the month of postpartum assessments and the prevalence of MDD and six instrument characteristics including the overall study quality rating; self-report versus provider reports; number of question items; the reference standard used; the diagnostic thresholds; and the type of diagnostic threshold (i.e. standard or optimal). A systematic pattern in a scatter plot was defined as a predictable variation in sensitivity values as values of specificity changed based on any of the investigated participant and/or instrument characteristics.

As per best practice guidelines, (Higgins and Thompson, 2002; Rutter and Gatsonis, 2001; Tosteson and Begg, 1988; Littenberg and Moses, 1993) meta-analyses were conducted only if three or more independent study samples with diagnostic performance values for the same instrument were available. The potential for study heterogeneity (i.e. more variation in instrument-specific sensitivity and specificity estimates than would be expected by chance alone) was assessed visually using scatter plots and forest plots. The small number of included studies and their respective sample sizes made the use of I^2 statistic and Cochran Q statistic tests for homogeneity not reliable, (Higgins and Thompson, 2002; Rutter and Gatsonis, 2001; Tosteson and Begg, 1988; Littenberg and Moses, 1993) and therefore these tests were not used.

2.6. Bayesian hierarchical summary receiver operating curve (HSROC) model

The pooled sensitivity and specificity of each case-finding instrument were generated using an adapted Bayesian hierarchical summary receiver operating (HSROC) model proposed by Dendukuri et al. (2012). This Bayesian HSROC model is an adaptation of the Rutter and Gatsonis (Rutter and Gatsonis, 2001) HSROC model that accounts for the variation in the *sensitivity* and *specificity* estimates of the same instrument due to the use of different diagnostic thresholds; the conditional dependence of errors generated from case-finding and reference standard results; the use of imperfect reference standards and; the use of different reference standards across studies. Technical details of this model are provided elsewhere (Dendukuri et al., 2012; Bernatsky et al., 2005; Dendukuri and Joseph, 2001).

Three sets of summary diagnostic performance estimates (i.e. antepartum, postpartum and the combined periods where possible) were generated for each case-finding instrument from two different Bayesian HSROC models.

Model A assumed conditional independence and the use of perfect reference standards. Conditional independence implies that conditional on the true MDD status of a participant, knowledge of the case-finding instrument result provides no information on the likelihood of the reference standard to be positive and vice versa.

Model B accounted for the conditional dependence and reference standard misclassification error. Conditional dependence implies that conditional on the true MDD status of a participant, knowledge of a case-finding instrument result influences the likelihood of the reference standard result to be positive (and vice versa) since they are both based on the same DSM-IV diagnostic classification criteria. Informative priors estimates of reference standard diagnostic performance used in Model B were based on results from expert panel validation studies (Ramirez Basco et al., 2000; Mitchell and Coyne, 2010; Miller et al., 2001).

Informative priors are a key part of Bayesian inference that represent information about an uncertain parameter (in our case the sensitivity and specificity estimates of reference standards) that is combined with the probability distribution of new data (TP, TN, FP, and FN) to yield a posterior distribution of pooled case-finding instrument sensitivity and specificity estimates from which summary estimates (i.e. median and 95% Bayesian Credible Intervals [BCI]) are estimated.

Reference standards examined for prior information included: Structured Clinical Interview of DSM Disorders (sensitivity range: 84–92%; specificity range: 91–98%), World Health Organization Composite International Diagnostic Interview (sensitivity range: 94–98%; specificity range: 72–79%), Schedule for Affective Disorders and Schizophrenia (sensitivity range: 74–84%; specificity range: 96–100%), and Diagnostic Interview Schedule (sensitivity range: 79–96%; specificity range: 90–98%). Two Bayesian HSROC

meta-analysis models (Models A and B) were implemented for each case-finding instrument and each peripartum period where possible using PROC MCMC in SAS (SAS Institute., 2009).

3. Results

3.1. Flow of included studies

A total of 1130 non-duplicated studies were identified through the search strategy, of which 70 (6%) were eligible for review in Phase II (Fig. 1). An additional four articles were identified from previous systematic reviews and moved to Phase II. Of the 74 articles reviewed in full in Phase II, 60 were excluded primarily due to the study of a non-eligible population (54 studies or 90%). Data on 21 MDD case-finding instruments reported in 14 eligible articles were retrieved in Phase III.

Data from the remaining 14 articles (19% of those read in full) containing sensitivity and specificity estimates of 21 different MDD case-finding instruments were extracted for potential meta-analysis. The diagnostic performance of six instruments (29% of the 21 instruments identified) namely the Edinburg Postnatal Depression Scale (EPDS10), (Ji et al., 2011; Tandon et al., 2012; Yonkers et al., 2009; Hanusa et al., 2008; Beck and Gable, 2001; Logsdon and Myers, 2010; O'Hara et al., 2012; Venkatesh et al., 2014; Chaudron et al., 2010) the Beck Depression Inventory version II (BDI-II), (Ji et al., 2011; Beck and Gable, 2001; O'Hara et al., 2012; Chaudron et al., 2010) the Center for Epidemiological Studies for Depression −20 items and Revised version (CESD20 and CESDR), (Tandon et al., 2012; Logsdon and Myers, 2010) the Patient Health Questionnaire (PHQ9), (Sidebottom et al., 2012; Hanusa et al., 2008; Davis et al., 2013; Gjerdingen et al., 2009) and two versions of the Hamilton Depression Rating Scales −17 and 21 items (HDRS17 and HDRS21) (Ji et al., 2011) was examined in at least three distinct study samples within and across peripartum periods. Two studies (Smith et al., 2010; Beck and Gable, 2005) were excluded from the meta-analyses because they did not include any of these six instruments.

3.2. Study and case-fining instrument characteristics

The study and case-finding instrument characteristics examined in the 12 studies included for meta-analysis are summarized in Table 1. Briefly, the study sample sizes ranged from 59 (Logsdon and Myers, 2010) to 1274 (Sidebottom et al., 2012) and MDD prevalence ranged from 1% (95%CI: 0%; 2%) (Yonkers et al., 2009) to 51% (95%CI: 36%; 66%) (Ji et al., 2011). The mean maternal age ranged from 16 years (standard deviation of 1) (Logsdon and Myers, 2010) to 33 years (standard deviation of 5) (Ji et al., 2011). Four of the six case-finding instruments were based on self-report assessments (EPDS10, CESD20/R, BDI-II and PHQ9) and two were provider-report assessments (HDRS17 and HDRS21). Two (CESD20/R and BDI-II) instruments had an easy literacy reading level (i.e. 3rd to 5th grade reading level) while the rest (EPDS10,

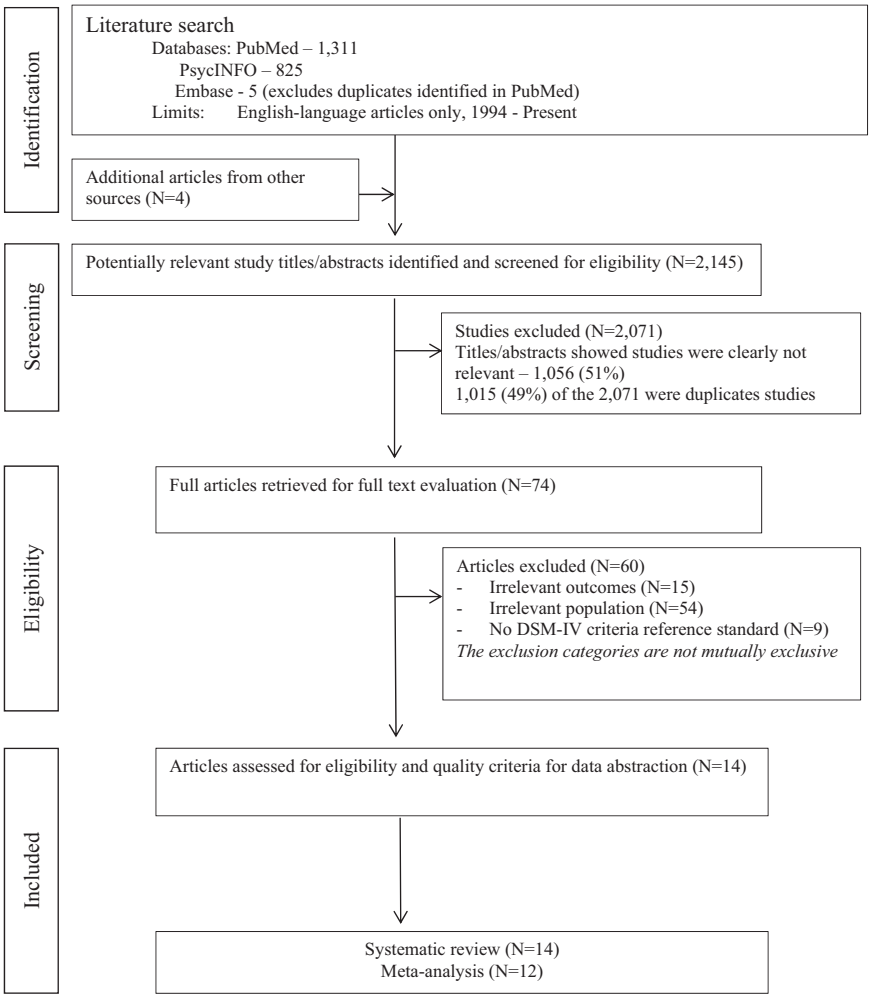

Fig. 1. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) flow diagram of literature search and study selection process.

Table 1
Summary characteristics of the 14 maternal MDD diagnostic accuracy studies included in the systematic review.

Author year	Participants' Characteristics			Study design (sample size)	Study setting	Reference standard	Case-finding instruments
	Percent multigravida	Race/ethnicity distribution	Age mean (SD)				
Ji et al., 2011	70%	W-86%, B-10%, H-3%	33(5)	Cohort (N = 534)	PC	SCID	HRSD17/21, BDI-II, EPDS10
Sidebottom et al., 2012	NR	W-10%, B-59%, H-8%	23(6)	Cross-sectional (N = 1274)	PC	SCID	PHQ9
Tandon et al., 2012	41%	B-100%	24(6)	Cross-sectional (N = 95)	CS/HV	SCID	CESDR, EPDS10, BDI-II
Yonkers et al., 2009	58%	W-80%, B-7%, H-10%	29(5)	Cohort (N = 838)	PC	CIDI	EPDS10
Hanusa et al., 2008	57%	W-72%, B-19%, H-NR	29(6)	Cross-sectional (N = 123)	PC	DIS	EPDS10, Brief PDSS, PHQ9
Beck and Gable, 2001	25%	W-87%, B-8%, H-4%	31(5)	Cross-sectional (N = 150)	CS	DIS	BDI-II, EPDS10, PDSS
Logsdon and Myers, 2010	0%	W-44%, B-42%, H-7%	16(1)	Cross-sectional (N = 59)	CS	KSADS-PL	EPDS10, CESD30, CESD20
Davis et al., 2013	NR	W-88%, B-5%, H-7%	29(5)	RCT (N = 1392)	PC	SCID	PHQ3/6/9, PRAMS3/6
O'Hara et al., 2012	57%	W-69%, B-10%, H-11%	27(5)	Cross-sectional (N = 1077)	PC & CS	SCID	EPDS2/3/7/10, PDSS, BDI-II, IDASGD, PRAMS2/3
Gjerdingen et al., 2009	58%	W-67%, B-18%, H-3%	29(6)	Cross-sectional (N = 506)	PC	SCID	PHQ2/9
Venkatesh et al., 2014	0%	W-16%, B-17%, H-53%	16(2)	RCT (N = 106)	PC	K-SCID	EPDS2/3/7/10
Chaudron et al., 2010	68%	W-17%, B-70%, H-7%	25(6)	Cross-sectional (N = 198)	CS	SCID	PDSS, BDI-II, EPDS10
Smith et al., 2010[a]	NR	W-63%, B-20%, H-10%	29(5)	Cross-sectional (N = 214)	PC	CIDI	PHQ2/8
Beck and Gable, 2005[a]	69%	H-100%	26(6)	Cross-sectional (N = 150)	CS	SCID	PDSS, PDSS-SF

Participants' characteristics: NR – Not Reported. B – African American, W – White, H – Hispanic.
Study setting: PC – Primary Care; CS – community setting; HV – Home visitation program.
SCID: Structured Clinical Interview of DSM Disorders; K-SCID – Kid's Version; SADS: Schedule for Affective Disorders and Schizophrenia; K-SADS PL: Kids Present and Lifetime version; CESD-R: Center of Epidemiological Studies-Depression Scale-Revised; BDI-II: Beck Depression Inventory version II; EPDS: Edinburg Postnatal Depression Scale; HDRS: Hamilton Depression Rating Scale.
[a] Studies excluded from the meta-analyses because they did not include any of the six commonly examined case-finding instruments across peripartum periods.

PHQ9, HDRS17 and HDRS21) had an average reading level (i.e. 6th to 9th reading level). Only four studies (Tandon et al., 2012; Logsdon and Myers, 2010; Beck and Gable, 2001; Chaudron et al., 2010) (33%) had a good overall rating of study quality; the rest (eight) had a fair study quality rating (Ji et al., 2011; Sidebottom et al., 2012; Yonkers et al., 2009; Hanusa et al., 2008; Davis et al., 2013; O'Hara et al., 2012; Gjerdingen et al., 2009; Venkatesh et al., 2014).

3.3. Assessment for heterogeneity

The scatter and forest plots suggested a considerable level of heterogeneity among studies. Therefore, Bayesian HRSOC models that account for between study heterogeneity were used to pool sensitivity and specificity estimates. Additionally, for two case-finding instruments examined in each peripartum-specific period (EPDS10 and BDI-II), there was a pattern of higher diagnostic performance during the antepartum than in the postpartum period. This suggested that the peripartum period modified the diagnostic performance results, and therefore, we examined study results within and across peripartum periods. Details of the analyses and results are provided elsewhere (Owora et al. 2016).

3.4. Comparisons between meta-analysis models across case-finding instruments

Compared to models adjusting for the conditional dependence of errors and reference standard misclassification error, the models assuming conditional independence and perfect reference standards systematically resulted in lower estimates of diagnostic performance for all six instruments (Table 2).

3.5. Diagnostic performance of the EPDS10

The EPDS10 was examined in eight studies with diagnostic thresholds ranging from 10 to 17 (Ji et al., 2011; Tandon et al., 2012; Yonkers et al., 2009; Hanusa et al., 2008; Beck and Gable, 2001; Logsdon and Myers, 2010; Venkatesh et al., 2014; Chaudron et al., 2010). Fig. 2 shows the HSROC plot of the EPDS10 sensitivity and specificity estimates based on 11 distinct study samples. Study-specific sensitivity and specificity estimates ranged from 63% to 94% and 83% to 90%, respectively. After adjusting for conditional dependence of errors and reference standard misclassification error, the pooled sensitivity and specificity were 82% (95%BCI: 50%; 98%) and 91% (95% BCI: 66%; 99%) respectively during the antepartum period. During the postpartum period, sensitivity was lower (74%; 95% BCI: 46%; 91%) but specificity was slightly higher (97%; 95% BCI: 84%; 99%).

3.6. Diagnostic performance of the BDI-II

Five studies examined the BDI-II across a range of diagnostic thresholds (12–20) (Ji et al., 2011; Tandon et al., 2012; Beck and Gable, 2001; O'Hara et al., 2012; Chaudron et al., 2010). Fig. 3 (HSROC plot) shows the study-specific sensitivity and specificity estimates that ranged from 55% to 92% and from 64% to 100%, respectively. After adjusting for conditional dependence of errors and reference standard misclassification error, the pooled sensitivity and specificity were 91% (95%BCI: 68%; 99%) and 89% (95% BCI: 62%; 98%) respectively during the antepartum period. The postpartum period estimates were less precise with sensitivity at 77% (95% BCI: 39%; 96%) and specificity at 93% (95% BCI: 53%; 99%).

3.7. Diagnostic performance of the CESD20/R

Fig. 4 shows the HSROC plot of the CESDR (Tandon et al., 2012)

Table 2
Hierarchical summary receiver operating curve pooled sensitivity and specificity results of commonly used MDD case-finding instruments among mothers of young children in the United States.

Instrument	Assessment period	Pooled sensitivity (95% BCI)		Pooled specificity (95% BCI)	
		Model A	Model B	Model A	Model B
CESD20/R	Across Peripartum	0.84 (0.61, 0.94)	0.90 (0.51, 0.99)	0.78 (0.48, 0.93)	0.80 (0.42, 0.98)
EPDS10	Antepartum	0.72 (0.59, 0.82)	0.82 (0.50, 0.98)	0.83(0.81, 0.86)	0.91 (0.66, 0.99)
EPDS10	Postpartum	0.68 (0.52, 0.81)	0.74(0.46, 0.91)	0.91 (0.81, 0.96)	0.97(0.84, 0.99)
EPDS10	Across Peripartum	0.67 (0.59, 0.77)	0.77 (0.54, 0.91)	0.89 (0.82, 0.94)	0.96 (0.87, 1.00)
BDI-II	Antepartum	0.84 (0.70, 0.93)	0.91 (0.68, 0.99)	0.81 (0.68, 0.90)	0.89 (0.62, 0.98)
BDI-II	Postpartum	0.69 (0.62, 0.76)	0.77(0.39, 0.96)	0.81 (0.77, 0.85)	0.93 (0.53, 0.99)
BDI-II	Across Peripartum	0.72 (0.67, 0.77)	0.86 (0.68, 0.97)	0.81 (0.78, 0.84)	0.92 (0.75, 0.98)
PHQ9	Across Peripartum	0.83 (0.78, 0.86)	0.92 (0.68, 0.99)	0.79 (0.64, 0.88)	0.79 (0.46, 0.91)
HDRS17	Antepartum	0.89 (0.82, 0.94)	0.97 (0.76, 1.00)	0.73 (0.67, 0.79)	0.78 (0.49, 0.94)
HDRS21	Antepartum	0.79 (0.70, 0.86)	0.85 (0.49, 0.98)	0.81 (0.75, 0.86)	0.89 (0.58, 0.99)

Model A: Bayesian HSROC effects model assuming conditional independence between case-finding instrument and reference standard test errors and perfect reference standards.
Model B: Bayesian HSROC Model adjusted for reference standard misclassification error and conditional dependence between case-finding instrument and reference standard test errors.

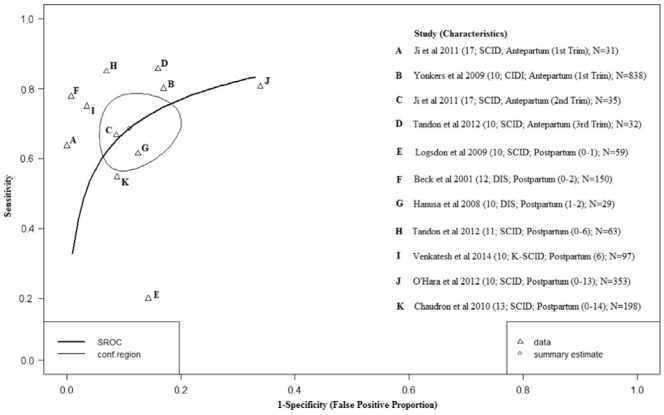

Fig. 2. Hierarchical Summary Receiver Operator Curve (HSROC) plot for the EPDS10. Each open triangle represents each study in the meta-analysis. The curve is the regression line that summarizes the overall diagnostic accuracy. The pooled sensitivity and specificity estimate is based on the assumption of conditional independence and use of perfect reference standards.

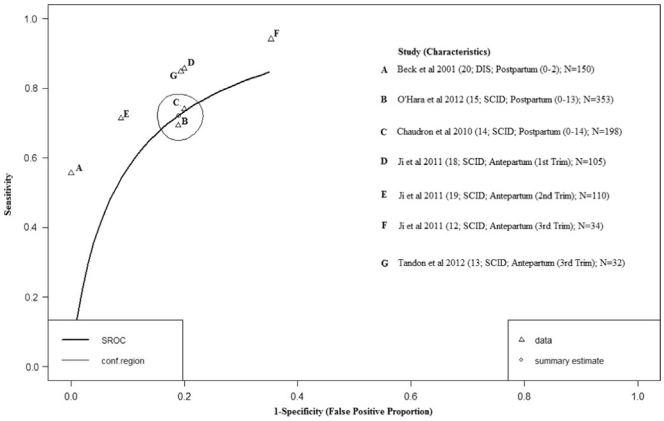

Fig. 3. Hierarchical Summary Receiver Operator Curve (HSROC) plot for the BDI-II. Each open triangle represents each study in the meta-analysis. The curve is the regression line that summarizes the overall diagnostic accuracy. The pooled sensitivity and specificity estimate is based on the assumption of conditional independence and use of perfect reference standards.

and CESD20 (Logsdon and Myers, 2010) sensitivity and specificity estimates based on three distinct study samples. One study examined diagnostic performance of the CESDR (diagnostic threshold: 21 and 22) in both the antepartum and postpartum periods

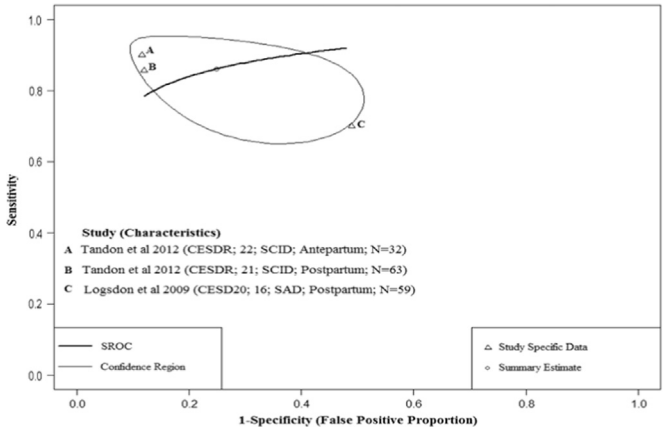

Fig. 4. Hierarchical Summary Receiver Operator Curve (HSROC) plot for the CESD20/R. Each open triangle represents each study in the meta-analysis. The curve is the regression line that summarizes the overall diagnostic accuracy. The pooled sensitivity and specificity estimate is based on the assumption of conditional independence and use of perfect reference standards.

while the other study assessed the CESD20 (diagnostic threshold: 16) during the postpartum period. Study-specific sensitivity estimates ranged from 68% to 88% and specificity ranged from 51% to 88%. Assuming perfect versus imperfect reference standards seemed to result in less uncertainty for the pooled specificity estimate but sensitivity estimates were not different (Table 2).

3.8. Diagnostic performance of the PHQ9

Four studies (Sidebottom et al., 2012; Hanusa et al., 2008; Davis et al., 2013; Gjerdingen et al., 2009) were included in the meta-analysis of PHQ9 (Fig. 5) at a diagnostic threshold of 10. Study-specific sensitivity estimates ranged from 32% to 85% and specificity ranged from 9% to 84%. After excluding Hanusa et al., (Hanusa et al., 2008) a study with a small sample size (29 participants) and outlier diagnostic performance estimates (i.e. observations that lie outside the general distribution of observed diagnostic performance estimates – see eTable 3 [greater than a 50% difference when compared to other study estimates of specificity]), there was overlap in the 95% BCIs of both the pooled sensitivity and specificity estimates when assuming perfect versus imperfect reference standards (Table 2).

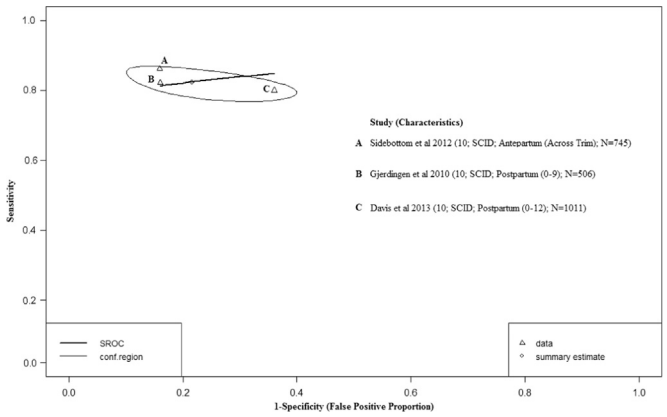

Fig. 5. Hierarchical Summary Receiver Operator Curve (HSROC) plot for the PHQ9. Each open triangle represents each study in the meta-analysis. The curve is the regression line that summarizes the overall diagnostic accuracy. The pooled sensitivity and specificity estimate is based on the assumption of conditional independence and use of perfect reference standards.

3.9. Diagnostic performance of the HDRS17 and HDRS21

One study provided sensitivity and specificity data for the HDRS17 and HDRS21 across a range of diagnostic thresholds (15 and 16–22 respectively) (Ji et al., 2011). There was no evidence of variation in sensitivity and specificity estimates across three trimesters as shown in Figs. 6a and b. HDRS17 sensitivity estimates ranged from 66% to 76% and specificity ranged from 78% to 88%. The HDRS21 had a slightly higher range of sensitivity estimates (70% to 81%) but similar specificity (78% to 88%). Additionally, no distributional differences were observed between summary diagnostic estimates generated under assumptions of perfect versus imperfect reference standards (Table 2).

3.10. Comparisons across peripartum periods (combined antepartum and postpartum period studies)

Comparisons across peripartum periods were possible for the EPDS10, BDI-II, CESD20/R, and PHQ9. After adjusting for the conditional dependence of errors and reference standard misclassification error, the PHQ9 had the highest pooled sensitivity estimates (92%; 95%BCI: 68%; 97%) followed by the CESD20/R (90%; 95% BCI: 51%; 99%), both the EPDS10 and BDI-II had pooled median estimates below 90%. The EPDS10 had the highest specificity estimate (96% ; 95%BCI: 87%; 100%) closely followed by BDI-II (92%; 95% BCI: 75%; 98%). Both the CESD20/R and PHQ9 had pooled median estimates at or below 80%.

4. Discussion

This is the first study to estimate the overall diagnostic performance of MDD case-finding instruments used among mothers of young children using a Bayesian Meta-analytical approach that accounts for varying diagnostic thresholds, use of multiple imperfect reference standards to validate the same case-finding instrument and the potential for conditional dependence of errors generated from case-finding instrument and reference standard results. In addition to these issues, previous attempts to carry out such analyses were limited by the lack of enough studies (i.e. less than three comparable studies) examining the same case-finding instrument (Gaynes et al., 2005; Agency for Healthcare Research and Quality AHRQ, 2014; Pignone et al., 2002; O'Connor et al., 2016; Gibson et al., 2009). As a consequence, these previous systematic reviews only provided a qualitative synthesis of the

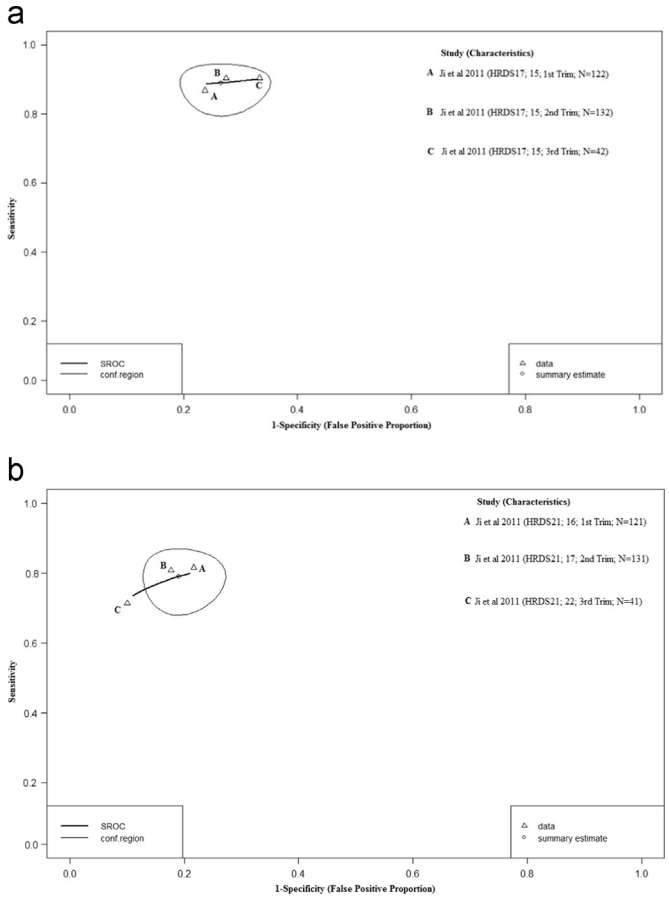

Fig. 6. a. Hierarchical Summary Receiver Operator Curve (HSROC) plot for the HDRS17. Each open triangle represents each study in the meta-analysis. The curve is the regression line that summarizes the overall diagnostic accuracy. The pooled sensitivity and specificity estimate is based on the assumption of conditional independence and use of perfect reference standards. b. Hierarchical Summary Receiver Operator Curve (HSROC) plot for the HDRS21. Each open triangle represents each study in the meta-analysis. The curve is the regression line that summarizes the overall diagnostic accuracy. The pooled sensitivity and specificity estimate is based on the assumption of conditional independence and use of perfect reference standards.

diagnostic performance of maternal MDD case-finding instruments (Gaynes et al., 2005; Agency for Healthcare Research and Quality AHRQ, 2014; Pignone et al., 2002; O'Connor et al., 2016; Gibson et al., 2009).

Our results suggest that maternal MDD case finding instruments have modest to high diagnostic performance across peripartum periods. MDD case finding instruments tended to show better sensitivity but worse specificity during the antepartum period while the opposite was true during the postpartum period. Results also show that failure to adjust for the above issues can substantially underestimate pooled sensitivity and specificity estimates and standard errors.

It is, therefore, plausible that previous meta-analysis estimates from general population diagnostic accuracy studies (Mulrow et al., 1995; Manea et al., 2015) were underestimated because they did not account for the methodological issues described above. For example, Mulrow et al., (Mulrow et al., 1995) using a linear random-effects model, found that the summary diagnostic performance of nine different case-finding instruments assessed in 18 primary-care based studies were modest at best with an overall sensitivity (84%; 95%CI: 79%; 89%) and specificity (72%; 95%CI: 67%; 77%). Using similar meta-analysis methods, pooled diagnostic performance estimates of seven studies showed that the PHQ9 had low sensitivity (55%; 95%CI: 39%; 73%) but high specificity (96%;

95%CI: 94%; 98%) (Manea et al., 2015). Clearly, these estimates are lower than those observed in our results before and after adjustment for conditional dependence of errors and reference standard misclassification error. Moreover, detection of MDD is expected to be more straight-forward in the general populations examined than among mothers of young children further supporting our suspicion of underestimated diagnostic performance results in previous meta-analysis studies.

Our meta-analysis represents a comprehensive evaluation of the diagnostic performance of MDD case-finding instruments used among mothers of young children in the US. Study selection, quality review and data extraction were conducted independently by two reviewers and disagreements were resolved by discussion/consensus. Quantitative analyses were performed in accordance with published guidelines (Rutter and Gatsonis, 2001; Tosteson and Begg, 1988; Littenberg and Moses, 1993; Deeks, 2001; Macaskill et al., 2010; Reitsma et al., 2009; Trikalinos and Balion, 2012). The Bayesian hierarchical models implemented for the summary of diagnostic performance adjusted for issues (described above) ignored in previous general population meta-analysis studies (Mulrow et al., 1995; Manea et al., 2015). Our HSROC models also accounted for the use of different reference standards with varying diagnostic performance across different diagnostic accuracy studies, a feature commonly encountered in the absence of gold standards for the diagnosis of health outcomes based on a patient's subjective experiences and perceptions of illness (Sadatsafavi et al., 2010; Chu et al., 2009; Walter et al., 1999; Dendukuri et al., 2012).

Despite these merits, this study has some limitations. For example, it was not possible to investigate how and to what extent patient characteristics (e.g. maternal age, race/ethnicity) and methodological issues (e.g. selection bias) may have affected study results due to the small number of studies examining each investigated instrument. As more MDD diagnostic accuracy studies among mothers of young children become available, meta-regression analyses that investigate the impact of these factors on summary diagnostic performance estimates are needed.

In summary, commonly used maternal MDD case-finding instruments in the US were found to have modest to high summary diagnostic performance that varied across instruments depending on the peripartum period of assessment. The variations in the summary diagnostic performance across instruments and peripartum specific periods point to the need for judicious decision making regarding which instruments should be used for maternal MDD case-finding depending on what tradeoffs may be acceptable as an opportunity cost (missing cases or incorrect diagnosis). Our results suggest the BDI-II and EPDS10 could be used to mitigate such costs in the antepartum and postpartum periods respectively. However, future research should quantify the costs associated with such tradeoffs to better inform decisions regarding the most cost-effective case-finding instruments in each specific peripartum period.

Author contributions

Arthur H. Owora had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: All authors.

Acquisition, analysis, and interpretation of the data: all authors.

Drafting of the manuscript: Arthur Owora.

Critical revision of the manuscript for important intellectual content: all authors.

All authors approved the final draft of the article.

Competing interests

None.

None of the contributing authors and I have any conflict of interest in this subject or any financial interest. The opinions expressed are those of the authors and do not necessarily reflect those of the Oklahoma University Health Sciences Center.

Funding

None.

Acknowledgments

The authors thank the Oklahoma University Health Sciences Bird Library Staff who provided invaluable assistance with literature search activities to identify publications relevant to this systematic review.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jad.2016.08.014.

References

Agency for Healthcare Research and Quality (AHRQ). Efficacy and Safety of Screening for Postpartum Depression. Comparative Effectiveness Review 106. Contract No. 290-2007-10066-I. ⟨http://www.ncbi.nlm.nih.gov/books/NBK137724⟩ (accessed 11.01.14).

Beck, C.T., Gable, R.K., 2001. Comparative analysis of the performance of the postpartum depression screening scale with two other depression instruments. Nurs. Res. 50 (4), 242–250.

Beck, C.T., Gable, R.K., 2005. Screening performance of the postpartum depression screening scale – Spanish version. J. Transcult. Nurs. 16 (4), 331–338.

Bernatsky, S., Joseph, L., Belisle, P., et al., 2005. Bayesian modelling of imperfect ascertainment methods in cancer studies. Stat. Med. 24 (15), 2365–2379.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., et al., 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann. Intern. Med. 138 (1), W1–12.

Chaudron, L.H., Szilagyi, P.G., Tang, W., et al., 2010. Accuracy of depression screening tools for identifying postpartum depression among urban mothers. Pediatrics 125 (3), e609–e617.

Chu, H., Chen, S., Louis, T.A., 2009. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. J. Am. Stat. Assoc. 104 (486), 512–523.

Davis, K., Pearlstein, T., Stuart, S., O'Hara, M., Zlotnick, C., 2013. Analysis of brief screening tools for the detection of postpartum depression: comparisons of the PRAMS 6-item instrument, PHQ-9, and structured interviews. Arch. Women's Ment. Health 16 (4), 271–277.

Deeks, J.J., 2001. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ 323 (7305), 157–162.

Dendukuri, N., Joseph, L., 2001. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 57 (1), 158–167.

Dendukuri, N., Schiller, I., Joseph, L., Pai, M., 2012. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. Biometrics 68 (4), 1285–1293.

Gaynes, B.N., Gavin, N., Meltzer-Brody, S., et al., 2005. Perinatal depression: prevalence, screening accuracy, and screening outcomes. Evid. Rep./Technol. Assess. 119, 1–8.

Gibson, J., McKenzie-McHarg, K., Shakespeare, J., Price, J., Gray, R., 2009. A systematic review of studies validating the edinburgh postnatal depression scale in antepartum and postpartum women. Acta Psychiat. Scand. 119 (5), 350–364.

Gjerdingen, D., Crow, S., McGovern, P., Miner, M., Center, B., 2009. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. Ann. Fam. Med. 7 (1), 63–70.

Hanusa, B.H., Scholle, S.H., Haskett, R.F., Spadaro, K., Wisner, K.L., 2008. Screening for depression in the postpartum period: a comparison of three instruments. J. Women's Health 17 (4), 585–596.

Higgins, J.P., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Stat. Med. 21 (11), 1539–1558.

Ji, S., Long, Q., Newport, D.J., et al., 2011. Validity of depression rating scales during pregnancy and the postpartum period: impact of trimester and parity. J.

Psychiatr. Res. 45 (2), 213–219.
Littenberg, B., Moses, L.E., 1993. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med. Decis. Mak. 13 (4), 313–321.
Logsdon, M.C., Myers, J.A., 2010. Comparative performance of two depression screening instruments in adolescent mothers. J. Women's Health 19 (6), 1123–1128.
Macaskill, P.G.C., Deeks, J.J., Harbord, R.M., Takwoingi, Y., 2010. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. the Cochrane Collaboration.
Manea, L., Gilbody, S., McMillan, D., 2015. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen. Hosp. Psychiatry 37 (1), 67–75.
Miller, P.R., Dasher, R., Collins, R., Griffiths, P., Brown, F., 2001. Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. Psychiatry Res. 105 (3), 255–264.
Mitchell, A.J., Coyne, J.C., 2010. Screening for Depression in Clinical Practice: An Evidence-Based Guide. Oxford University Press, New York, USA.
Mulrow, C.D., Williams Jr., J.W., Gerety, M.B., Ramirez, G., Montiel, O.M., Kerber, C., 1995. Case-finding instruments for depression in primary care settings. Ann. Intern. Med. 122 (12), 913–921.
O'Connor, E., Rossom, R.C., Henninger, M., Groom, H.C., Burda, B.U., 2016. 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 315 (4), 388–406.
O'Hara, M.W., Stuart, S., Watson, D., Dietz, P.M., Farr, S.L., D'Angelo, D., 2012. Brief scales to detect postpartum depression and anxiety symptoms. J. Women's Health 21 (12), 1237–1243.
Owora, HA., Helene, C., Reese, J., Garwe, T., 2016. Diagnostic performance of major depression disorder case-finding instruments used among mothers of young children in the United States: A systematic review. J Affect Disord 201, 185–193.
Pereira, A.T., Marques, M., Soares, M.J., et al., 2014. Profile of depressive symptoms in women in the perinatal and outside the perinatal period: similar or not? J. Affect. Disord. 166, 71–78.
Philipp Doebler. Mada: Meta-Analysis of Diagnostic Accuracy. R package version 0.5.7. ⟨http://CRAN.R-project.org/package=mada⟩ (accessed 30.01.15).
Pignone, M.P., Gaynes, B.N., Rushton, J.L., et al., 2002. Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. Ann. Intern. Med. 136 (10), 765–776.
R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from ⟨http://www.R-project.org⟩ (accessed 30.01.15).
Ramirez Basco, M., Bostic, J.Q., Davies, D., et al., 2000. Methods to improve diagnostic accuracy in a community mental health setting. Am. J. Psychiatry 157 (10), 1599–1605.
Reitsma, J.B., Rutjes, A.W.S., Whiting, P., Vlassov, V.V., Leeflang, M.M.G., Deeks, J.J., 2009. Assessing methodological quality (Chapter 9). Deeks, J.J., Bossuyt, P.M., Gatsonis, C. (Eds.), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. the Cochrane Collaboration.
Rutter, C.M., Gatsonis, C.A., 2001. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat. Med. 20 (19), 2865–2884.
Sadatsafavi, M., Shahidi, N., Marra, F., et al., 2010. A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard, combining random-effect and latent-class methods to estimate test accuracy. J. Clin. Epidemiol. 63 (3), 257–269.
Sidebottom, A.C., Harrison, P.A., Godecker, A., Kim, H., 2012. Validation of the patient health questionnaire (PHQ)−9 for prenatal depression screening. Arch. Women's Ment. Health 15 (5), 367–374.
Smith, M.V., Gotman, N., Lin, H., Yonkers, K.A., 2010. Do the PHQ-8 and the PHQ-2 accurately screen for depressive disorders in a sample of pregnant women? Gen. Hosp. Psychiatry 32 (5), 544–548.
Tandon, S.D., Cluxton-Keller, F., Leis, J., Le, H.N., Perry, D.F., 2012. A comparison of three screening tools to identify perinatal depression among low-income African American women. J. Affect. Disord. 136 (1–2), 155–162.
Tosteson, A.N.A., Begg, C.B., 1988. A General regression methodology for ROC curve estimation. Med. Decis. Mak. 8 (3), 204–215.
Trikalinos, T.A., Balion, C.M., 2012. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". J. Gen. Intern. Med. 27 (Suppl. 1), S67–S75.
Venkatesh, K.K., Zlotnick, C., Triche, E.W., Ware, C., Phipps, M.G., 2014. Accuracy of brief screening tools for identifying postpartum depression among adolescent mothers. Pediatrics 133 (1), e45–e53.
Walter, S.D., Irwig, L., Glasziou, P.P., 1999. Meta-analysis of diagnostic tests with imperfect reference standards. J. Clin. Epidemiol. 52 (10), 943–951.
Whiting, P.F., Rutjes, A.W., Westwood, M.E., et al., 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155 (8), 529–536.
Yonkers, K.A., Smith, M.V., Gotman, N., Belanger, K., 2009. Typical somatic symptoms of pregnancy and their impact on a diagnosis of major depressive disorder. Gen. Hosp. Psychiatry 31 (4), 327–333.