Published in final edited form as:

J Affect Disord. 2016 September 01; 201: 185–193. doi:10.1016/j.jad.2016.05.015.

Diagnostic performance of major depression disorder case-finding instruments used among mothers of young children in the United States: A systematic review

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

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

Abstract

Introduction—Growing recognition of the interrelated negative outcomes associated with major depression disorder (MDD) among mothers and their children has led to renewed public health interest in the early identification and treatment of maternal MDD. Healthcare providers, however, remain unsure of the validity of existing case-finding instruments. We conducted a systematic review to identify the most valid maternal MDD case-finding instrument used in the United States.

Methods—We identified articles reporting the sensitivity and specificity of MDD case-finding instruments based on *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) by systematically searching through three electronic bibliographic databases, PubMed, PsycINFO, and EMBASE, from 1994 to 2014. 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.

Results—Overall, we retrieved 996 unduplicated articles and selected 74 for full-text review. Of these, 14 articles examining 21 different instruments were included in the systematic review. The 10 item Edinburgh Postnatal Depression Scale and Postpartum Depression Screening Scale had the most stable (lowest variation) and highest diagnostic performance during the antepartum and postpartum periods (sensitivity range: 0.63–0.94 and 0.67–0.95; specificity range: 0.83–0.98 and 0.68–0.97 respectively). Greater variation in diagnostic performance was observed among studies with higher MDD prevalence.

*Correspondence to: 745 Martina Lane, Edmond, OK 73034, United States.

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 of the contributing authors have any conflict of interest 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.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.jad.2016.05.015>.

Limitation—Factors that explain greater variation in instrument diagnostic performance in study populations with higher MDD prevalence were not examined.

Discussion—Findings suggest that the diagnostic performance of maternal MDD case-finding instruments is peripartum period-specific.

Keywords

Diagnostic performance; Case-finding instrument; Maternal depression; Major depression disorder

1. Background

Major depression disorder (MDD) is a mental health condition characterized by the presence of five or more symptoms that reflect a change in mood or impaired functioning (weight change, insomnia, irritability, loss of energy, worthlessness, diminished concentration and suicidality) including a depressed mood or anhedonia (loss of interest or pleasure) for at least a two week period (American Psychiatric Association, 2000).

In the United States (US), the proportion of mothers with mental health and behavioral issues who get screened for MDD is low due to poor healthcare insurance coverage for such service, lack of treatment training and low prioritization for screening among primary care providers (Horwitz et al., 2007). The low prioritization of screening has been in part due to the lack of empirical evidence on the cost-effectiveness of screening and early treatment on longer term maternal-child outcomes which is underscored by inconsistent recommendations by various stakeholders (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) (Agency for Healthcare Research and Quality (AHRQ), 2014; Gaynes et al., 2005; Siu, 2016). Nonetheless, the United States Preventive Services Task Force recently recommended that all pregnant and postpartum mothers be screened for maternal MDD, showing that this issue is increasingly becoming a concern for stakeholders (Siu, 2016).

The overall societal and economic costs associated with untreated maternal MDD are likely to be considerable since MDD does not only affect the mother, but also her children (Ammerman et al., 2010; Grupp-Phelan and Ammerman, 2010; Lovejoy, 1991; Surkan et al., 2014, 2012). Depressed women have been shown to have poorer health, psychosocial outcomes and lower quality of life than non-depressed women (Farr et al., 2011; Ertel et al., 2011). Children of depressed mothers suffer from higher levels of difficulty in general functioning, attachment problems, and mood disorders than children of non-depressed mothers (Goodman et al., 2011; Beardslee et al., 1998). These problems among children are also associated with maternal stress, a known risk factor of maternal MDD (Pereira et al., 2014; Marcus and Heringhausen, 2009; Brown and Solchany, 2004; Ammerman et al., 2010). Since the costs and burden associated with these interrelated negative outcomes is substantial, it is essential to determine how well instruments developed to measure maternal MDD perform.

The clinical diagnosis MDD is not objective because it is in part based on subjective patient experiences and perceptions. As a result of its subjective nature, high prevalence and the distress it causes, researchers and clinicians have developed a large number of maternal MDD diagnostic instruments for screening, case-finding, and diagnosis as well as for monitoring patients' progress through the course of treatment (Agency for Healthcare Research and Quality (AHRQ), 2014; Gaynes et al., 2005). Despite the existence of numerous maternal MDD diagnostic instruments, the few studies that have evaluated their diagnostic validity have shown important variability of results within and among case-finding instruments, making it difficult to choose any particular one for clinical use. In addition, summarizing diagnostic performance of MDD screening and case-finding instruments is challenging because maternal MDD depends on the presence of symptoms that may be perceived and reported differently across cultures and peripartum periods (Horwitz et al., 2007; Pereira et al., 2014). Indeed, MDD may be under or over-diagnosed due to the presence symptoms that mimic those of true MDD especially during the immediate postpartum period (Pereira et al., 2014). There is also heterogeneity among existing maternal MDD diagnostic accuracy studies in terms of settings, study populations, choice of diagnostic thresholds and reference standards used to validate case-finding instrument results (Agency for Healthcare Research and Quality (AHRQ), 2014; Gaynes et al., 2005). In the systematic review conducted by Gaynes et al. (2005), 23 maternal MDD diagnostic accuracy studies published between 1980 and 2004 examined four case-finding instruments: Edinburgh Postnatal Depression Scale, Postpartum Depression Screening Scale, Beck Depression Inventory and the Center for Epidemiological Studies on Depression Scale. An updated review conducted by Myers et al. (2013) included 18 postpartum studies published between 2004 and 2012 that examined three additional instruments: Patient Health Questionnaire, Antenatal Risk Questionnaire and the Hamilton Rating Scale for Depression (Agency for Healthcare Research and Quality (AHRQ), 2014). Of the 41 reviewed studies (Agency for Healthcare Research and Quality (AHRQ), 2014; Gaynes et al., 2005), only eight were conducted in the US. Clearly, this limits the generalizability of overall review findings to US mothers because the cultures and health systems in the other countries differ from that of the US. Additionally, the mixture of mild and severe MDD symptoms may affect the diagnostic performance of case-finding instruments across different countries and study populations. In light of these limitations and gaps in previous systematic reviews, the objective of this study is to identify the most valid MDD case finding-instruments used among mothers of young children in the United States.

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, 2014. An experienced search 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. Details of the search strategy are summarized in eTable 1 (a–c). Briefly, our search strategy included various terms for maternal depression, diagnostic performance, and the names of existing

published MDD case-finding instruments. 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 by two investigators Jessica Reese (JR) and AO to determine their eligibility using criteria summarized in eTable 2a. Briefly, studies measuring MDD among mothers of young children (0–5 years old) in the US and reporting of both case-finding and reference standard instrument results met the inclusion criteria. Articles that included mothers from other countries or mothers of older children (> 5 years) were excluded. Eligible 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 the four quality criteria domains based on Quality Assessment of Diagnostic Accuracy Studies - Second Version (QUADAS-2) guidelines: (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 guidelines) were added to assess the potential for confounding and effect modification of the relationship between case-finding instrument and reference standard results. Here, confounding refers to the distortion of the relationship between case-finding instrument and reference standard results due to a third variable (e.g. peripartum period of assessment), whereas effect modification refers to a change in that relationship depending on the levels or categories of the third variable (e.g. antepartum versus postpartum). Each 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 2b and 2c, respectively. A study with a low risk of bias classification for all four QUADAS-2 domains was assigned a 'good' overall quality rating. Studies with a low risk of bias for three or two and one or none of the QUADAS-2 domains were assigned a '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

All analyses were conducted in R 3.1.2 (R Development Core, 2010) using the Meta-analysis of Diagnostic Accuracy - MADA package (R Development Core Team, 2015; Doebler, 2015). Sensitivity, specificity, and 95% confidence intervals for each diagnostic threshold and period of assessment (i.e. antepartum and postpartum) for the case-finding instruments described in the studies included from Phase III were estimated.

Sensitivity was defined as the proportion of positive tests among mothers with MDD while specificity was the proportion of negative tests among mothers without MDD as determined by reference standard diagnostic interviews. All reference standards were assumed to perfectly classify maternal MDD. The diagnostic thresholds used to determine positive and negative case-finding instrument results were based on either (a) recommendations from existing instrument-specific development and/or psychometric validation studies (i.e. *standard diagnostic thresholds*) or (b) maximizing the summation of sensitivity and specificity in specific study samples using 'gold standard' or reference standard diagnostic interviews (i.e. *optimal diagnostic thresholds*). Only DSM-IV compliant reference standard results were examined. The reference standards included: Structured Clinical Interview of DSM Disorders (SCID), World Health Organization Composite International Diagnostic Interview (CIDI), Schedule for Affective Disorders and Schizophrenia (SADS), and Diagnostic Interview Schedule (DIS). 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 (American Psychiatric Association, 2000). Postpartum depression was defined as an episode of MDD with the onset of symptoms occurring after childbirth (range: 1–14 months).

Systematic patterns in the scatter plots of sensitivity and specificity (1-false positive proportion) estimates were examined according to 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 (i.e. non-random) variation in sensitivity values as values of specificity changed based on any of the investigated participant and/or instrument characteristics.

3. Results

3.1. Flow of included studies

A total of 996 non-duplicated studies were identified through the search strategy, of which 69 (7%) were eligible for review in Phase II (Fig. 1). An additional five 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 (Ji et al., 2011; Sidebottom et al., 2012; Yonkers et al., 2009; Tandon et al., 2012; Smith et al., 2010; Venkatesh et al., 2014; Beck and Gable, 2001, 2005; Chaudron et al., 2010; Gjerdingen et al., 2009; Davis et al., 2013; Hanusa et al., 2008; Logsdon and Myers, 2010; O'Hara et al., 2012).

3.2. Quality assessments of the 14 included studies

A summary of the evaluation of each signaling question can be found in eFigure 1. None of the studies examined the potential for confounding or effect modification by study participant characteristics on case-finding instrument diagnostic performance. Five (36%) studies (Tandon et al., 2012; Beck and Gable, 2001, 2005; Chaudron et al., 2010; Logsdon and Myers, 2010) had a good overall quality rating and the remaining nine studies had a fair quality rating. There were no systematic differences in the diagnostic performance of instruments based on overall quality of comparable studies (i.e. studies with similar participant characteristics). Based on the quality assessment, all 14 studies were deemed acceptable for inclusion in the quantitative evaluation.

3.3. Study population characteristics of the 14 included studies

Participant and study characteristics as well as the list of the case-finding instruments and reference standard examined in the 14 included studies are summarized in Table 1. Twelve studies (80%) included mixed race/ethnicity populations. The mean maternal age ranged from 16 to 33 years. MDD prevalence ranged from 1–45% based on reference standard assessments.

3.4. Case-finding instrument characteristics

The characteristics of the 21 case-finding instruments examined in the 14 included studies are summarized in Table 2. The Edinburgh Postnatal Depression Scale (EPDS10) was the most commonly assessed case-finding instrument while the Structured Clinical Interview of DSM-IV Disorders (SCID) was the most used reference standard (nine studies for each; 64% each). Nineteen (90%) of the case-finding instruments examined were based on self-report. The number of question items ranged from two (Patient Health Questionnaire - PHQ2 and EPDS2) to 35 (Postpartum Depression Screening Scale - PDSS) items. Seven (33%) of the case-finding instruments examined had an easy literacy reading level (i.e. 3rd to 5th grade reading level) while the rest (14) had an average reading level (i.e. 6th to 9th reading level).

3.5. Diagnostic performance of the MDD case-finding instruments at different peripartum periods

The diagnostic performance is defined by the sensitivity and the specificity of each MDD case finding instrument to correctly classify a mother as suffering or not suffering from MDD, respectively.

For case-finding instruments examined in each peripartum-specific period (Edinburgh Postnatal Depression Scale - EPDS10, Beck's Depression Inventory second version - BDI-II, Center for Epidemiological Studies Depression Scale Revised - CESDR, Patient Health Questionnaire - PHQ9, Hamilton Depression Rating Scale - HDRS17 and HDRS21), there was a pattern of higher diagnostic performance (i.e. sensitivity and specificity estimates) during the antepartum period as compared to the postpartum period. This suggested that the peripartum period modified the diagnostic performance results, and therefore, these periods were not combined in further analyses.

For the antepartum period (eTable 3), four case-finding instruments (CESDR (Tandon et al., 2012), EPDS10 (Ji et al., 2011; Yonkers et al., 2009; Tandon et al., 2012), BDI-II (Ji et al., 2011; Tandon et al., 2012), PHQ9 (Sidebottom et al., 2012), and HDRS17 (Ji et al., 2011), HDRS21 (Ji et al., 2011)) were assessed across three pregnancy trimesters (1st, 2nd and 3rd) (Ji et al., 2011; Sidebottom et al., 2012; Yonkers et al., 2009; Tandon et al., 2012). Across different studies and instruments examined, the sensitivity estimates ranged from 0.63 (95%CI: 0.35 – 0.84) (Ji et al., 2011) in the first trimester to 0.94 (95%CI: 0.73–0.99) (Yonkers et al., 2009) in the second trimester for the EPDS10 while the specificity estimates ranged from 0.64 (95%CI: 0.41 – 0.82) (Ji et al., 2011) in the third trimester for the BDI-II to 0.98 (95%CI: 0.81 – 1.00) (Ji et al., 2011) in the first trimester for the EPDS10. There was a pattern of higher diagnostic performance during the second and third trimesters than during the first trimester across different studies and instruments. The EPDS10 had the least variation and highest diagnostic performance across different trimesters and diagnostic thresholds (sensitivity range: 0.63–0.94 and specificity range: 0.83 – 0.98). Compared to self-report based case-finding instruments (e.g. BDI-II and EPDS10), provider-report based case-finding instruments (HDRS17 and HDRS21) had less variable diagnostic performance across trimesters (1st, 2nd and 3rd).

For the postpartum period (eTable 4), ten studies (Tandon et al., 2012; Beck and Gable, 2001, 2005; Chaudron et al., 2010; Gjerdingen et al., 2009, 2011; Davis et al., 2013; Hanusa et al., 2008; Logsdon and Myers, 2010; O'Hara et al., 2012) examined 18 case-finding instruments (BDI-II, (Tandon et al., 2012; Beck and Gable, 2001; Chaudron et al., 2010; O'Hara et al., 2012) CESD20, (Logsdon and Myers, 2010) CESD30, (Logsdon and Myers, 2010) CESDR, (Tandon et al., 2012) EPDS [2, 3, 7, 10], (Tandon et al., 2012; Beck and Gable, 2001; Chaudron et al., 2010; Hanusa et al., 2008; Logsdon and Myers, 2010; O'Hara et al., 2012) Inventory of Depression and Anxiety Symptoms - IDAS, (O'Hara et al., 2012) PDSS, (Beck and Gable, 2001, 2005; Chaudron et al., 2010) PDSS-Short Form, (Hanusa et al., 2008; Beck and Gable, 2005) PHQ2 Likert Scale, (Gjerdingen et al., 2009) PHQ2 Nominal Scale, (Gjerdingen et al., 2009) PHQ9, (Gjerdingen et al., 2009, 2011; Davis et al., 2013; Hanusa et al., 2008) Pregnancy Risk Assessment Questionnaire - PRAMS [2, 3, 6] (Davis et al., 2013; O'Hara et al., 2012)). The sensitivity estimates ranged from 0.23

(95%CI: 0.07 – 0.52) (Logsdon and Myers, 2010) for the EPDS10 to 1.00 (95% CI: 0.99 – 1.00) (Davis et al., 2013) for the PRAM6 while the specificity estimates ranged from 0.09 (95%CI: 0.02 – 0.31) (Hanusa et al., 2008) to 1.00 (95%CI: 0.99 – 1.00) (Davis et al., 2013) for the PHQ9. There was a pattern of lower and more variable diagnostic performance during the first two months postpartum than later. The PDSS (diagnostic threshold/cut-point scores: 50–80) showed least variation and highest diagnostic performance across different postpartum periods and diagnostic thresholds (sensitivity range: 0.67–0.95 and specificity range: 0.68–0.97).

For both the antepartum and postpartum periods, there was a pattern of lower case-finding instrument diagnostic performance with increasing MDD prevalence. Fig. 2 shows the variation in the diagnostic performance of two commonly used case-finding instruments (EPDS10 and BDI-II) by peripartum period and MDD prevalence. There were no systematic patterns in the variation of diagnostic performance values by the overall study quality rating, reference standard used, diagnostic thresholds used for the same instrument and the types of diagnostic threshold used (i.e. optimal versus standard thresholds). For some of the instruments that had versions with different numbers of question items (e.g. EPDS [10, 7, 3 and 2], PHQ [9, 8, 4 and 2], PRAMS [6, 3 and 2], PDSS [35 and 7], and HRDS [21 and 17]), there was a pattern of lower specificity but higher sensitivity with fewer question items in comparable study samples (i.e. same study setting, peripartum period, trimester or postpartum month of assessment, maternal age, and MDD prevalence).

4. Discussion

Our findings suggest that the diagnostic performance of maternal MDD case-finding instruments is peripartum period-specific and that the amount of variation in diagnostic performance depends on the underlying prevalence of MDD. Across the 14 different studies that examined 21 instruments, diagnostic performance estimates tended to be higher during the antepartum than postpartum period. The EPDS10 and PDSS had the most stable and highest diagnostic performance during the antepartum and postpartum periods respectively.

The peripartum period during which MDD is assessed impacted the diagnostic performance of some, but not all instruments. In some instances, the lack of variation could be explained by the small period-specific study sample sizes.

Some of the systematic patterns or variation in diagnostic performance observed among case-finding instruments examined may be explained by either the complex pathophysiology of depressive symptoms among mothers, spectrum bias or social desirability bias.

The pathophysiology of MDD among mothers is complex because some of the depressive symptoms included in the definition of MDD are considered normal responses during the first few days after childbirth, a condition often referred to as the "baby blues" Hirst and Moutier (2010). During this period, there may be over-diagnosis of MDD illness because the symptoms of "baby blues" and stress associated with sleep deprivation, adapting to the feeding schedule of infants, need for close supervision or attention and care of medically fragile children may be confused for MDD Liu and Alloy (2010). Although none of the 14

reviewed studies adjusted for the presence of “baby blues” or stress-related factors, lower sensitivity and specificity values were generally observed when MDD assessments were conducted during the first month postpartum versus later months using the same case-finding instrument, suggesting that such bias may be present.

Another explanation for variation of the diagnostic performance across peripartum periods is the presence of spectrum bias. Spectrum bias generally occurs when a study only includes (1) “easy to diagnose” or “definite” cases not representing those with milder symptoms, artificially improving the sensitivity, or (2) “easy to rule out” or non-cases not representing those with symptoms that could be confused as MDD symptoms, artificially improving the specificity Mulherin and Miller (2002), (Goehring et al., 2004). Although the spectrum of participants cannot be directly measured, studies conducted in populations with a higher prevalence of MDD are more likely to represent higher surveillance for MDD. As a result, MDD cases in such populations might include mothers with fewer or milder symptoms, leading to lower case-finding instrument sensitivity estimates. Our results show a pattern of lower sensitivity estimates in studies conducted in populations with higher prevalence of MDD than those with lower prevalence supporting the likely presence of spectrum bias. Previous systematic reviews that examined the diagnostic performance of MDD case-finding instruments among mothers failed to explore the potential impact of such bias in their synthesis of results across diagnostic accuracy studies Agency for Healthcare Research and Quality (AHRQ) (2014), (Gaynes et al., 2005). In fact, two different case definitions (minor and major MDD) were often used interchangeably to classify MDD, further constraining the ability to validly evaluate data across studies. Agency for Healthcare Research and Quality (AHRQ) (2014), (Gaynes et al., 2005) From a quantitative perspective, accounting for spectrum bias effects will require either stratification or statistical adjustments for the underlying prevalence of MDD in future meta-analyses Mower (1999), Mulherin and Miller (2002).

The final possible explanation for variation of the diagnostic performance across peripartum periods is social desirability bias. This bias occurs if a mother over-reports MDD-related symptoms or severity to elicit empathy or sympathy for their circumstances or environmental and social support. Such bias is likely to be more prevalent during the postpartum than the antepartum period when maternal stress is likely to be present. Pereira et al. (2014), (Ammerman et al., 2010) Based on our results, the larger amount of variation in sensitivity and specificity estimates among studies that assessed postpartum versus antepartum mothers using comparable instruments (i.e. diagnostic threshold and reference standards) is probably indicative of such social desirability bias effects.

Our review also noted some variation between self-report and provider-report case-finding instruments. This too may be due to social desirability bias since self-reported MDD would be expected to be influenced by such bias more than provider-reported MDD. Unfortunately, only two of the reviewed instruments in our study (HDRS17 and HDRS21) were based on provider-reports, making it difficult to assess if such bias is linked to type of reporter. Nonetheless, these provider-report instruments tended to have higher and less variable diagnostic performance across trimesters when compared to the EPDS10 and BDI-II.

Another source of variability in diagnostic performance is race/ethnicity through another form of social desirability bias. Mothers with different race/ethnicities may have different cultural perceptions of MDD illness which could affect how they respond to question items. However, the reviewed studies did not report race/ethnic group specific estimates of instrument diagnostic performance, making it impossible to qualitatively evaluate the impact of social desirability bias in this context.

In order to attenuate the effects of potential bias (especially false positives due to low specificity) due to the "baby blues" in the immediate post-partum period, spectrum bias or social desirability bias, our results suggest the use of case-finding instrument versions with more than fewer question items may be ideal. However, consistent with the recommendations in the studies that examined instrument versions with different question items, (Ji et al., 2011; Smith et al., 2010; Venkatesh et al., 2014; Gjerdingen et al., 2009; Davis et al., 2013; Logsdon and Myers, 2010; O'Hara et al., 2012; Beck and Gable, 2005) the importance of instruments with fewer question items as initial screeners in relatively busy clinical settings where lengthy instruments may not be feasible should not be overlooked.

4.1. Limitations

None of the reviewed studies examined the diagnostic performance of MDD case-finding instruments mothers with children aged between two and five years, a period of time considered critical to a child's development. Cummings and Davies (1994); Heckman (2006) There is clearly a need for such studies given the potential negative effect that maternal MDD could have on both mothers and their children. Additionally, because of the small number of studies examining each investigated instrument, 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. As more maternal MDD diagnostic accuracy studies become available, the impact of these factors on case-finding instrument diagnostic performance needs to be empirically investigated. Further research is also needed to identify factors that may explain greater variation in instrument diagnostic performance estimates in study populations with higher MDD prevalence to better inform clinical practice recommendations.

5. Conclusion

In summary, this systematic review represents a comprehensive qualitative evaluation of the diagnostic performance of MDD case-finding instruments used among mothers of young children in the US. Study findings have shown that the diagnostic performance of maternal MDD case-finding instruments vary across peripartum periods and also depend on the underlying prevalence of MDD in the study populations of interest. In determining which case-finding instruments to use, healthcare providers and public health professionals should carefully consider the period of assessment as well as the study population characteristics.

Acknowledgments

Funding

None.

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.

References

- American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4. Washington, DC: 2000. Text Revision
- Agency for Healthcare Research and Quality (AHRQ). [Accessed January 11, 2014] 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>)
- Gaynes BN, Gavin N, Meltzer-Brody S, et al. Perinatal depression: prevalence, screening accuracy, and screening outcomes. *Evid. Rep./Technol. Assess.* 2005; 119:1–8.
- Siu AL. US Preventive Services Task Force (USPSTF), 2016. Screening for depression in adults: us preventive services task force recommendation statement. *J. Am. Med. Assoc.* 2016; 315(4):380–387.
- Horwitz SM, Kelleher KJ, Stein RE, et al. Barriers to the identification and management of psychosocial issues in children and maternal depression. *Pediatrics.* 2007; 119(1):208–218.
- Pereira AT, Marques M, Soares MJ, et al. Profile of depressive symptoms in women in the perinatal and outside the perinatal period: similar or not? *J. Affect Disord.* 2014; 166:71–78. [PubMed: 25012412]
- Marcus SM, Heringhausen JE. Depression in childbearing women: when depression complicates pregnancy. *Prim. Care.* 2009; 36(1):151–159. [PubMed: 19231607]
- Brown MA, Solchany JE. Two overlooked mood disorders in women: subsyndromal depression and prenatal depression. *Nurs. Clin. North Am.* 2004; 39(1):83–95. [PubMed: 15062729]
- Ammerman RT, Putnam FW, Bosse NR, Teeters AR, Van Ginkel JB. Maternal depression in home visitation: a systematic review. *Aggress. Violent Behav.* 2010; 15(3):191–200. [PubMed: 20401324]
- Grupp-Phelan J, Ammerman RT. Maternal depression and child growth: definitional issues, longitudinal trajectories, and analytic considerations. *J. Pediatr.* 2010; 157(3):359–360. [PubMed: 20542289]
- Lovejoy MC. Maternal depression: effects on social cognition and behavior in parent-child interactions. *J. Abnorm Child. Psychol.* 1991; 19(6):693–706. [PubMed: 1791274]
- Surkan PJ, Ettinger AK, Hock RS, Ahmed S, Strobino DM, Minkovitz CS. Early maternal depressive symptoms and child growth trajectories: a longitudinal analysis of a nationally representative US birth cohort. *BMC Pediatr.* 2014; 14:185. [PubMed: 25047367]
- Surkan PJ, Ettinger AK, Ahmed S, Minkovitz CS, Strobino D. Impact of maternal depressive symptoms on growth of preschool- and school-aged children. *Pediatrics.* 2012; 130(4):847–855.
- Farr SL, Hayes DK, Bitsko RH, Bansil P, Dietz PM. Depression, diabetes, and chronic disease risk factors among us women of reproductive age. *Prev. Chronic Dis.* 2011; 8(6):A119. [PubMed: 22005612]
- Ertel KA, Rich-Edwards JW, Koenen KC. Maternal depression in the United States: nationally representative rates and risks. *J. Womens Health.* 2011; 20(11):1609–1617.
- Goodman SH, Rouse MH, Connell AM, Broth MR, Hall CM, Heyward D. Maternal depression and child psychopathology: a meta-analytic review. *Clin. Child. Fam. Psychol. Rev.* 2011; 14(1):1–27. [PubMed: 21052833]
- Beardslee WR, Versage EM, Gladstone TR. Children of affectively ill parents: a review of the past 10 years. *J. Am. Acad. Child. Adolesc. Psychiatry.* 1998; 37(11):1134–1141. [PubMed: 9808924]
- Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *J. Am. Med. Assoc.* 1999; 282(11):1061–1066.
- Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern Med.* 2011; 155(8):529–536. [PubMed: 22007046]

- Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann. Intern Med.* 2003; 138(1):W1–12. [PubMed: 12513067]
- 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 January 30, 2015]
- Doebler, Philipp. [Accessed January 30, 2015] Mada: Meta-Analysis of Diagnostic Accuracy. R package version 0.5.7. <http://CRAN.R-project.org/package=mada>
- Ji S, Long Q, Newport DJ, et al. Validity of depression rating scales during pregnancy and the postpartum period: impact of trimester and parity. *J. Psychiatr. Res.* 2011; 45(2):213–219. [PubMed: 20542520]
- Sidebottom AC, Harrison PA, Godecker A, Kim H. Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. *Arch. Women's Ment. Health.* 2012; 15(5):367–374. [PubMed: 22983357]
- Yonkers KA, Smith MV, Gotman N, Belanger K. Typical somatic symptoms of pregnancy and their impact on a diagnosis of major depressive disorder. *Gen. Hosp. Psychiatry.* 2009; 31(4):327–333. [PubMed: 19555792]
- Tandon SD, Cluxton-Keller F, Leis J, Le HN, Perry DF. A comparison of three screening tools to identify perinatal depression among low-income African American women. *J. Affect Disord.* 2012; 136(1–2):155–162. [PubMed: 21864914]
- Smith MV, Gotman N, Lin H, Yonkers KA. Do the PHQ-8 and the PHQ-2 accurately screen for depressive disorders in a sample of pregnant women? *Gen. Hosp. Psychiatry.* 2010; 32(5):544–548. [PubMed: 20851275]
- Venkatesh KK, Zlotnick C, Triche EW, Ware C, Phipps MG. Accuracy of brief screening tools for identifying postpartum depression among adolescent mothers. *Pediatrics.* 2014; 133(1):e45–e53. [PubMed: 24344102]
- Beck CT, Gable RK. Comparative analysis of the performance of the Postpartum Depression Screening Scale with two other depression instruments. *Nurs. Res.* 2001; 50(4):242–250. [PubMed: 11480533]
- Chaudron LH, Szilagyi PG, Tang W, et al. Accuracy of depression screening tools for identifying postpartum depression among urban mothers. *Pediatrics.* 2010; 125(3):e609–e617. [PubMed: 20156899]
- Gjerdingen D, Crow S, McGovern P, Miner M, Center B. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann. Fam. Med.* 2009; 7(1):63–70. [PubMed: 19139451]
- Gjerdingen D, McGovern P, Center B. Problems with a diagnostic depression interview in a postpartum depression trial. *J. Am. Board Fam. Med.* 2011; 24(2):187–193. [PubMed: 21383219]
- Davis K, Pearlstein T, Stuart S, O'Hara M, Zlotnick C. 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.* 2013; 16(4):271–277. [PubMed: 23579244]
- Hanusa BH, Scholle SH, Haskett RF, Spadaro K, Wisner KL. Screening for depression in the postpartum period: a comparison of three instruments. *J. Womens Health (Larchmt.).* 2008; 17(4):585–596. [PubMed: 18345995]
- Logsdon MC, Myers JA. Comparative performance of two depression screening instruments in adolescent mothers. *J. Womens Health (Larchmt.).* 2010; 19(6):1123–1128. [PubMed: 20500127]
- O'Hara MW, Stuart S, Watson D, Dietz PM, Farr SL, D'Angelo D. Brief scales to detect postpartum depression and anxiety symptoms. *J. Womens Health (Larchmt.).* 2012; 21(12):1237–1243. [PubMed: 23130750]
- Beck CT, Gable RK. Screening performance of the postpartum depression screening scale–Spanish version. *J. Transcult. Nurs.* 2005; 16(4):331–338. [PubMed: 16160195]
- Mower WR. Evaluating bias and variability in diagnostic test reports. *Ann. Emerg. Med.* 1999; 33(1):85–91. [PubMed: 9867892]
- Hirst KP, Moutier CY. Postpartum major depression. *Am. Fam. Physician.* 2010; 82(8):926–933. [PubMed: 20949886]

- Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann. Intern Med.* 2002; 137(7):598–602. [PubMed: 12353947]
- Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. *Stat. Med.* 2004; 23(1):125–135. [PubMed: 14695644]
- Liu RT, Alloy LB. Stress generation in depression: A systematic review of the empirical literature and recommendations for future study. *Clin. Psychol. Rev.* 2010; 30(5):582–593. [PubMed: 20478648]
- Cummings EM, Davies PT. Maternal depression and child development. *J. Child. Psychol. Psychiatry.* 1994; 35(1):73–112. [PubMed: 8163630]
- Heckman JJ. Skill formation and the economics of investing in disadvantaged children. *Science.* 2006; 312(5782):1900–1902. [PubMed: 16809525]
- Streiff LD. Can clients understand our instructions? *Image J. Nurs. Sch.* 1986; 18(2):48–52. [PubMed: 3634751]

Fig. 1.

PRISMA flow diagram of literature search and study selection process.

Legend: PRISMA -preferred reporting items for systematic reviews and meta-analyses.

Fig. 2.

Scatter plot of the sensitivity (y-axis) and 1-specificity (x-axis) estimates from EPDS10 and BDI-II diagnostic accuracy studies. Legend: Study specific sensitivity and specificity estimates during the antepartum (circle), postpartum (triangle) and combined (square) periods. Shapes inside the symbols are a size scale gradient indicating prevalence estimates from low to high. Diagnostic threshold range: EPDS10 (8–17) and BDI-II (11–20).

Legend: Study specific sensitivity and specificity estimates during the antepartum (circle), postpartum (triangle) and combined (square) periods. Shapes inside the symbols are a size scale gradient indicating prevalence estimates from low to high. Diagnostic threshold range: EPDS10 [8–17] and BDI-II [11–20].

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				
Ji et al. (2011)	70%	W-86%, B-10%, H-3%	Cohort (N = 534)	PC	SCID	HRSD17/21, BDI-II, EPDS10
Sidebottom et al. (2012)	NR	W-10%, B-59%, H-8%	Cross-sectional (N = 1274)	PC	SCID	PHQ9
Tandon et al. (2012)	41%	B-100%	Cross-sectional (N = 95)	CS/HV	SCID	CESDR, EPDS10, BDI-II
Yonkers et al. (2009)	58%	W-80%, B-7%, H-10%	Cohort (N = 838)	PC	CIDI	EPDS10
Hanusa et al. (2008)	57%	W-72%, B-19%, H-NR	Cross-sectional (N = 123)	PC	DIS	EPDS10, PDSS-SF, PHQ9
Beck and Gable (2005)	69%	H-100%	Cross-sectional (N = 150)	CS	SCID	PDSS, PDSS-SF
Beck and Gable (2001)	25%	W-87%, B-8%, H-4%	Cross-sectional (N = 150)	CS	DSM-IV	BDI-II, EPDS10, PDSS
Logsdon and Myers (2010)	0%	W-44%, B-42%, H-7%	Cross-sectional (N = 59)	CS	KSADS-PL	EPDS10, CESD20/30
Davis et al. (2013)	NR	W-88%, B-5%, H-7%	RCT (N = 1392)	PC	SCID	PHQ3/6/9, PRAMS3/6
O'Hara et al. (2012)	57%	W-69%, B-10%, H-11%	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%	Cross-sectional (N = 506)	PC	SCID	PHQ2/9
Venkatesh et al. (2014)	0%	W-16%, B-17%, H-53%	RCT (N = 106)	PC	K-SCID	EPDS2/3/7/10
Chaudron et al. (2010)	68%	W-17%, B-70%, H-7%	Cross-sectional (N = 198)	CS	SCID	PDSS, BDI-II, EPDS10
Smith et al. (2010)	NR	W-63%, B-20%, H-10%	Cross-sectional (N = 214)	PC	CIDI	PHQ2/8

Participants' characteristics: NR - Not Reported. B-Black or 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 SCID Version; CESD-R: Center of Epidemiological Studies-Depression Scale-Revised; BDI-II: Beck Depression Inventory version II; EPDS: Edinburgh Postnatal Depression Scale; HDRS: Hamilton Depression Rating Scale; PDSS: Postnatal Depression Screening Scale; SF: Short Form; PHQ: Patient Health Questionnaire; PRAMS: Pregnancy Risk Assessment Monitoring System Survey

Table 2

Summary of the 21 MDD case-finding instruments evaluated in the systematic review (14 studies).

ID	Instrument	Year ^C	Instrument Version	Type	Time (mins)	Reference Period	No. items	Score Range	Standard Cut-Point	Literacy Level
1	BDI (Ji et al., 2011; Tandon et al., 2012; Beck and Gable, 2001; Chaudron et al., 2010; O'Hara et al., 2012)	1996	BDI-II	SR	2–5	Previous 2 weeks	21	0–63	20	Easy
2	CESD (Tandon et al., 2012; Logsdon and Myers, 2010)	1977	CESD20	SR	5–10	Past week	20	0–60	16	Easy
3		1977	CESD30	SR	5–10	Past week	30	0–90	24	Easy
4		2004	CESDR	SR	5–10	Previous 2 weeks	20	0–60	16	Easy
5	PDSS (Beck and Gable, 2001; Chaudron et al., 2010; Hanusa et al., 2008; Beck and Gable, 2005)	2000	PDSS	SR	5–10	Previous 2 weeks	35	35–175	80	Average
6		2002	Brief PDSS	SR	5–10	Previous 2 weeks	7	7 to –35	14	Average
7	EPDS (Ji et al., 2011; Yonkers et al., 2009; Tandon et al., 2012; Venkatesh et al., 2014; Beck and Gable, 2001) (Chaudron et al., 2010; Hanusa et al., 2008; Logsdon and Myers, 2010; O'Hara et al., 2012)	1987	EPDS10	SR	2–5	Past week	10	0–30	13	Average
8		2008	EPDS7	SR	< 2	Past week	10	0–30	13	Average
9		2008	EPDS3	SR	< 2	Past week	10	0–30	13	Average
10		2008	EPDS2 IS	SR	< 2	Past week	10	0–30	13	Average
11		2008	EPDS2 RS	SR	< 2	Past week	3	0–6	3	Average
12	PHQ (Sidebottom et al., 2012; Smith et al., 2010; Gjerdingen et al., 2009; Davis et al., 2013; Sidebottom et al., 2012)	2001	PHQ9	SR	2–5	Previous 2 weeks	9	0–27	10	Average
13		2002	PHQ8	SR	2–5	Previous 2 weeks	8	0–24	10	Average
14		2010	PHQ4	SR	2–5	Previous 2 weeks	4	0–12	10	Average
15		2010	PHQ2	SR	2–5	Previous 2 weeks	2	0–6	3	Average
16	PRAMS (Davis et al., 2013; O'Hara et al., 2012)	2012	PRAMS6	SR	2–5	Since delivery	6	0–30	15	Easy
17		2012	PRAMS3	SR	2–5	Since delivery	3	0–15	9	Easy
18		2012	PRAMS2	SR	2–5	Since delivery	3	0–15	9	Easy
19	HDRS (Ji et al., 2011)	1960	HDRS17	PR	15–20	Past week	17	0–54	18	Average
20		1967	HDRS21	PR	15–20	Past week	21	0–54	18	Average
21	IDAS (O'Hara et al., 2012)	2007	IDAS-GD	SR	5–10	Previous 2 weeks	20	20–100	54	Average

Report: Mode of assessment (SR- Self-report; PR – Provider/Clinician report); IS- Inflated Scores; RS- Raw Scores

Year^C: Year instrument was created/validated

Standard Cut-point: diagnostic threshold (symptom cut-off scores) recommended by instrument developers

Time (minutes): Expected self-report or interview completion time

Literacy levels: Easy – 3rd to 5th grade reading level; Average – 6th to 9th grade reading level according to Fog formula (Streff, 1986)

CESDR: Center of Epidemiological Studies-Depression Scale-Revised (this instrument has a two week reference period)

CESD20: Center of Epidemiological Studies-Depression Scale-Revised (this instrument has a one week reference period)