Draft Comparative Effectiveness Review

Number XX

Efficacy and Safety of Screening for Postpartum Depression

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

This information is distributed solely for the purposes of predissemination peer review. It has not been formally disseminated by the Agency for Healthcare Research and Quality (AHRQ). It does not represent and should not be construed to represent the Agency for Healthcare Research and Quality or the Department of Health and Human Services determination or policy.

Contract No.

Prepared by:

Investigators:

AHRQ Publication No. xx-EHCxxx
<Month Year>

Preface

[To be provided in the final report]

Acknowledgments

[To be provided in the final report.]

Key Informants

[To be provided in the final report.]

Technical Expert Panel

[To be provided in the final report.]

Peer Reviewers

[To be provided in the final report.]

Efficacy and Safety of Screening for Postpartum Depression

Structured Abstract

Objectives: To describe the benefits and harms of specific screening tools and strategies for screening for postpartum depression.

Data Sources: We searched PubMed®, Embase®, PsycINFO®, and the Cochrane Database of Systematic Reviews for relevant English-language studies published from January 1, 2004, through January 25, 2012, that evaluated the performance of screening instruments for postpartum depression, potential benefits and harms of screening, and impact on appropriate post-screening actions.

Review Methods: Two investigators screened each abstract and full-text article for inclusion; abstracted data; and performed quality ratings, applicability ratings, and evidence grading. A simulation model was used to estimate the effects of screening for postpartum depression on the overall balance of benefits and harms.

Results: Fifteen studies of nine different screening instruments were identified—nine for the Edinburgh Postnatal Depression Scale (EDPS), four for the Postpartum Depression Screening Scale (PDSS), two for different versions of the Beck Depression Inventory (BDI), and one each of the other instruments. Heterogeneity in setting, patient population, and choice of threshold prevented formal synthesis. For most tests in most studies, sensitivity and specificity were in the 80–90 percent range, with higher sensitivity associated with lower specificity. Confidence intervals for sensitivity and specificity broadly overlapped. Fourteen studies presented results for risk factors for postpartum depression. Adverse pregnancy outcomes and chronic medical conditions (low strength of evidence) and past history of mood disorder, poor relationship quality, and poor social support were all associated with an increased risk of postpartum depression (moderate strength of evidence). Based on two studies, there was insufficient evidence to evaluate whether timing relative to delivery, setting, or provider affected test characteristics of screening instruments. Based on four studies, there was low strength of evidence that screening resulted in decreased depressive symptoms and improved mental health, but an increased number of unscheduled infant doctor visits in women who were screened compared with women who were not screened. Based on two studies, referral rates for women with positive screening tests were highest when screening was done during the delivery admission (low strength of evidence). Modeling suggests serial testing may be a reasonable strategy to reduce false positives while minimizing false negatives.

Conclusions: Although currently available screening instruments are reasonably sensitive and specific in detecting postpartum depression, there is insufficient evidence to draw any conclusions about the net balance of benefits and harms of screening for postpartum depression, or about whether specific tools or strategies would result in a more favorable balance.

Contents

Executive Summary **ES-1**
Introduction 1
Background 1
Postpartum Depression 1
Screening for Postpartum Depression 2
Scope and Key Questions 6
Scope of the Review 6
Key Questions 7
Analytic Framework 8
Methods 11
Topic Refinement and Review Protocol 11
Literature Search Strategy 11
Search Strategy 11
Inclusion and Exclusion Criteria 12
Study Selection 15
Data Extraction 15
Quality Assessment of Individual Studies 16
Data Synthesis 17
Strength of the Body of Evidence 19
Applicability 20
Peer Review and Public Commentary 20
Results 21
Introduction 21
Results of Literature Searches 21
Description of Included Studies 22
Key Question 1. Performance Characteristics of Screening Instruments 23
Key Points 23
Description of Included Studies 24
Detailed Synthesis 24
Key Question 2. Effect of Individual Subject Factors on Screening Performance 32
Key Points 32
Description of Included Studies 32
Detailed Synthesis 33
Key Question 3. Effect of Testing Variables on Screening Performance 40
Key Points 40
Description of Included Studies 40
Detailed Synthesis 41
Key Question 4. Comparative Benefits of Screening 42
Key Points 42
Description of Included Studies 42
Detailed Synthesis 42
Key Question 5. Comparative Harms of Screening 44
Key Points 44
Description of Included Study 45

Detailed Synthesis 45
Key Question 6. Factors Affecting the Likelihood of an Appropriate Action After a Positive Screening Result 45
Key Points 45
Description of Included Studies 45
Detailed Synthesis 46
Discussion 49
Key Findings and Strength of Evidence 49
KQ 1. Performance Characteristics of Screening Instruments 49
KQ 2. Effect of Individual Subject Factors on Screening Performance 52
KQ 3. Effect of Testing Variables on Screening Performance 54
KQ 4. Comparative Benefits of Screening; and KQ 5. Comparative Harms of Screening 54
KQ 6. Factors Affecting the Likelihood of an Appropriate Action After a Positive Screening Result 55
Findings in Relationship to What is Already Known 56
Applicability 57
Implications for Clinical and Policy Decisionmaking 57
Potential Value of Simulation Modeling 58
Limitations of the Comparative Effectiveness Review Process 61
Limitations of the Evidence Base 62
Research Gaps 63
General Gaps 63
KQ 1 63
KQ 2 64
KQs 3–6 64
Conclusions 65
References 66
Abbreviations 73

Tables

Table 1. DSM-IV-TR diagnostic criteria for major depressive disorder 1
Table 2. Guidelines/recommendations for screening for postpartum depression 3
Table 3. Inclusion and exclusion criteria 12
Table 4. Definitions of overall quality ratings 17
Table 5. Strength of evidence—required domains 19
Table 6. Test characteristics for postpartum depression screening instruments other than EPDS or PDSS 28
Table 7. Maternal demographic risk factors for postpartum major depression 33
Table 8. Obstetric history risk factors for postpartum depression 35
Table 9. Sensitivity and specificity of screening instruments by timing of screening and gravidity 36
Table 10. General medical history risk factors for postpartum depression 37
Table 11. Psychiatric history risk factors for postpartum major depression 37
Table 12. Relationship and social support risk factors for postpartum major depression 38
Table 13. Sensitivity and specificity of screening instruments in the early and late postpartum periods 41

Table 14. Strength of evidence domains for test characteristics of screening tests for postpartum depression 50
Table 15. Effect of prevalence of major depression on true positives, false positives, and false negatives at varying levels of sensitivity and specificity 52
Table 16. Strength of evidence domains for associations with patient characteristics and risk of postpartum depression 53
Table 17. Strength of evidence domains for the effect of varying timing, setting, and provider on screening for postpartum depression 54
Table 18. Strength of evidence domains for benefits and harms of screening for postpartum depression 55
Table 19. Strength of evidence domains for the effect of timing, setting, and provider on rates of referral and treatment among women with a positive screening test for postpartum depression 56

Figures

Figure 1. Analytic framework 9
Figure 2. Literature flow diagram 22
Figure 3. Sensitivity of the EPDS at various thresholds 25
Figure 4. Specificity of the EPDS at various thresholds 26
Figure 5. Sensitivity of the PDSS at various thresholds 27
Figure 6. Specificity of the PDSS at various thresholds 28
Figure 7. Comparative sensitivity of various screening instruments 30
Figure 8. Comparative specificity of various screening instruments 31
Figure 9. Estimated false negatives and false positives for each of nine tests as either a single screen (“One Test”) or as part of a serial testing strategy following a positive response to one of the two questions (“Two Test”) at a prevalence of 5.8% 59
Figure 10. Acceptabilty curve for trade-off between false postives (“costs”) and treated depression (“effectiveness”) at different thresholds for false positives/treated depression ratio (“willingness-to-pay”). 60

Appendixes

Appendix A. Exact Search Strings
Appendix B. Data Abstraction Elements
Appendix C. List of Included Studies
Appendix D. List of Excluded Studies
Appendix E. Study Characteristics Table

Effective Health Care

Efficacy and Safety of Screening for Postpartum Depression

Executive Summary

The Effective Health Care Program was initiated in 2005 to provide valid evidence about the comparative effectiveness of different medical interventions. The object is to help consumers, health care providers, and others in making informed choices among treatment alternatives. Through its Comparative Effectiveness Reviews, the program supports systematic appraisals of existing scientific evidence regarding treatments for high-priority health conditions. It also promotes and generates new scientific evidence by identifying gaps in existing scientific evidence and supporting new research. The program puts special emphasis on translating findings into a variety of useful formats for different stakeholders including consumers.

The full report and this summary are available at www.effectivehealthcare.ahrq.gov/reports/final.cfm

Background

Condition and Preventive Strategies

Depression is a potentially life-threatening condition with a substantial impact on quality of life. The impact of depression in postpartum women is at least as great as that for depression in other populations. Postpartum depression is defined in the *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision* (hereafter, DSM-IV-TR)[1] as a major depressive disorder according to standard diagnostic criteria, namely 5 or more of the following symptoms present during the same 2-week period, with a secondary criterion of onset of symptoms within 4 weeks of delivery:

a. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful);
b. Markedly diminished interest in pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others);
c. Significant weight loss when not dieting or weight gain (e.g., change of more than 5% body weight in a month), or decrease or increase in appetite nearly every day;
d. Insomnia or hypersomnia nearly every day;
e. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down);
f. Fatigue or loss of energy nearly every day;
g. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick);
h. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either subjective account or as observed by others);

i. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide).

A new set of diagnostic criteria for psychiatric illness, the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (DSM-V), is currently scheduled for release in May 2013; preliminary discussions suggest that the overall diagnostic framework for postpartum depression (i.e., major depression with a specification of postpartum onset) will remain unchanged, although the window for diagnosis may be extended to 6 months after delivery.

Other diagnostic standards allow the definition of onset to extend beyond 4 weeks and up to 12 months after delivery and/or add a "minor depression" subcategory (2 to 4 of the symptoms listed above). There is high-quality evidence for effective treatment of patients who meet criteria for major depression in other settings; evidence is inconsistent for postpartum depression.[2-4]

The most recent U.S.–based synthesis of the evidence, performed for the Agency for Healthcare Research and Quality (AHRQ) in 2005,[2,3] estimated that the prevalence of major depression alone during the first postpartum year is 1.0 to 5.9 percent, and that the prevalence of major and minor depression combined is 6.5 to 12.9 percent. Incidence estimates for the first 3 postpartum months were up to 6.5 percent for major depression alone and 14.5 percent for major and minor depression.

In addition to the direct effect of postpartum depression on maternal quality of life and the risk of suicide, maternal postpartum depression has also been associated with an increased risk of infant mortality, adverse effects on some measures of infant development, and increased health care resource utilization for both mothers and infants.

Given the potential impact of postpartum depression on maternal and infant health, there has been considerable interest in strategies aimed at identifying women who are at risk for postpartum depression or who have postpartum depression, with the ultimate goal being the application of effective preventive or therapeutic interventions. All major organizations providing care to pregnant and postpartum women and infants recognize the risk of postpartum depression and the potential benefit of screening, but the strength of recommendations is variable. For example, no U.S.–based organizations recommend use of a specific screening instrument. Factors limiting the strength of recommendations include the lack of sufficient data on the most appropriate screening instrument, the optimal time(s) for screening, issues concerning reimbursement and the scope of practice, and the need for adequate systems for ensuring appropriate care for women identified through screening. In addition to uncertainty about the benefits of screening for postpartum depression, there is almost no evidence on potential harms; given that many of the signs and symptoms included in the diagnostic criteria for depression are common and normal responses to pregnancy, childbirth, and caring for infants, the risk of false positive results could potentially be relatively high. In addition, many studies include the diagnostic category of minor depression, despite a lack of evidence for effective interventions for symptoms that do not meet criteria for a diagnosis of depression.

There is consistent uncertainty about how well currently available tests and strategies for identifying women with, or at risk for, postpartum depression perform, and about how factors such as timing relative to delivery, setting, and provider might affect the performance of these strategies, and the factors that influence effective management of positive results. In addition, there is a paucity of evidence on the overall balance of harms and benefits of screening for postpartum depression compared with no screening, or between different strategies.

Scope and Key Questions

This Comparative Effectiveness Review (CER) was funded by AHRQ and is designed to evaluate the comparative diagnostic accuracy, benefits, and harms of available screening instruments for postpartum depression. Further details are provided under "Key Questions" and "Analytic Framework," below, and in the section on "Inclusion and Exclusion Criteria" in the Methods chapter. For each Key Question (KQ), we further consider whether the diagnostic accuracy, benefits, and harms of the screening instruments evaluated differ among specific patient subgroups of interest, including subgroups defined by age, race, parity, history of mood disorders, history of intimate partner violence, perinatal outcomes, or cultural factors, or by the timing, setting, or provider.

By summarizing data that support improved screening of postpartum depression, we hope to enhance patient-centered outcomes for women, their partners, and children, and to reduce health care utilization and costs. This project will benefit patients, providers, payers, and policymakers. Patients will benefit from more robust data on the comparative diagnostic accuracy, benefits, and harms of available screening instruments for postpartum depression. Providers will benefit by gaining a better understanding of which tools, timing, and settings are most effective for screening women for postpartum depression, and which patients benefit most from such screening. Policymakers will be able to design and implement programs to make better use of scarce health care resources while improving the health status of women at risk for postpartum depression.

The Key Questions (KQs) considered in this CER are:

KQ 1: This question has two parts:

a. What are the sensitivity and specificity of currently available screening instruments for detecting postpartum depression, and how do these translate into the likelihood of false-negative and false-positive results in different populations and settings?
b. Are there clinically relevant differences in the ability of currently available screening instruments to correctly identify specific signs or symptoms of depression (e.g., suicidal ideation)?

KQ 2: This question has two parts:

a. Are there individual factors (age, race, parity [number of live births], history of mood disorders, history of intimate partner violence, perinatal outcomes, cultural factors) that affect the baseline risk of postpartum depression and, therefore, the subsequent positive and negative predictive values of screening instruments?
b. Are there validated predictive models or algorithms based on such factors that would improve the performance of screening instruments?

KQ 3: Are the performance characteristics (sensitivity, specificity, predictive values) of screening instruments affected by:

a. Timing (prenatal, peripartum, or at various times in the first postpartum year) and frequency of screening?
b. Setting (prenatal visit, hospital/birthing center/home, postpartum maternal visit, or well-child visit)?
c. Provider (obstetrician, midwife, pediatrician, family practitioner, other health provider)?

KQ 4: What are the comparative benefits of screening for postpartum depression when compared to no screening, or between different screening strategies (based on choice of screening instrument, timing, setting, etc.)?

KQ 5: What are the comparative harms of screening for postpartum depression when compared to no screening, or between different screening strategies (based on choice of screening instrument, timing, setting, etc.)?

KQ 6: Is the likelihood of an appropriate action (referral, diagnosis, treatment, etc.) after a positive screening result affected by timing, setting, patient characteristics, or other factors?

Methods

The methods for this Comparative Effectiveness Review (CER) follow those suggested in the Agency for Healthcare Research and Quality (AHRQ) *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (hereafter referred to as the *Methods Guide*)[5] and *Methods Guide for Medical Test Reviews* (hereafter referred to as the *Medical Test Guide*).[6]

Input from Stakeholders

During the topic refinement stage, we solicited input to help define the KQs from Key Informants representing medical professional societies/clinicians in the areas of mental health, obstetrics and gynecology, women's health, pregnancy and perinatal epidemiology, psychiatry, maternal and fetal medicine, pediatrics, and primary care; patients, scientific experts; and payers. The KQs were then posted for public comment for 4 weeks from November 8 to December 6, 2011, and the comments received were considered in the development of the research protocol. We next convened a Technical Expert Panel (TEP) comprising clinical, content, and methodological experts to provide input in defining populations, interventions, comparisons, and outcomes, and in identifying particular studies or databases to search. The Key Informants and members of the TEP were required to disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts. Any potential conflicts of interest were balanced or mitigated. Neither Key Informants nor members of the TEP performed analysis of any kind, nor did any of them contribute to the writing of this report. Members of the TEP were invited to provide feedback on an initial draft of the review protocol which was then refined based on their input, reviewed by AHRQ, and posted for public access at the AHRQ Effective Health Care Website.[7]

Literature Search Strategy

To identify the relevant published literature, we searched PubMed®, Embase®, PsycINFO®, and the Cochrane Database of Systematic Reviews (CDSR), limiting the search to studies published from January 1, 2004, to January 25, 2012 (subsequent to the March 2004 search end date of the 2005 AHRQ evidence report on postpartum depression).[2,3] Where possible, we used existing validated search filters (such as the Clinical Queries Filters in PubMed). An experienced search librarian guided all searches. We supplemented the electronic searches with a manual search of references from a set of key primary and systematic review articles. All citations were imported into an electronic database (EndNote® X4; Thomson Reuters, Philadelphia, PA).

We used several approaches to identify relevant grey literature. These included searches of trial registry and conference abstract databases for relevant articles from completed studies and requests to publishers of proprietary depression screening tools for scientific information packets. Grey literature databases included ClinicalTrials.gov; the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) search portal; and ProQuest COS Conference Papers Index.

As a mechanism to ascertain publication bias, we searched ClinicalTrials.gov to identify completed but unpublished studies. While the draft report is under peer review, we will update the literature searches and include any eligible studies identified either during that search or through peer or public reviews in the final report.

Inclusion and Exclusion Criteria

Criteria used to screen articles for inclusion/exclusion at both the title-and-abstract and full-text screening stages are detailed in Table 3 of the main report. For all KQs, the search focused on studies that were conducted in economically developed countries, published since 2004 in English-language journals, and reported screening instrument performance characteristics or the effects of screening for postpartum depression in a population of pregnant women or women during the first 12 months after delivery. The following outcomes were considered: screening instrument performance characteristics; diagnosis of depression; receipt of appropriate diagnostic and treatment services for symptoms of depression; scores on validated measures of maternal well-being and parenting; breastfeeding; scores on validated diagnostic instruments for depression; health-related quality of life; maternal suicidal or infanticidal behaviors; scores on validated instruments of infant health and development; maternal and infant health system resource utilization; and scores on validated measures of stigmatization. Studies providing data for fathers or domestic partners were also considered; outcomes assessed for this group included scores on validated mental health instruments, health-related quality of life, and health system resource utilization.

Study Selection

Using the prespecified inclusion and exclusion criteria, titles and abstracts were reviewed independently by two investigators for potential relevance to the KQs. Articles included by either reviewer underwent full-text screening. At the full-text review stage, paired researchers independently reviewed the articles and indicated a decision to “include” or “exclude” the article for data abstraction. When the two reviewers arrived at different decisions about whether to include or exclude an article, they reconciled the difference through review and discussion, or through a third-party arbitrator if needed. Full-text articles meeting our eligibility criteria were included for data abstraction. Relevant review articles, meta-analyses, and methods articles were flagged for manual searching of references and cross-referencing against the library of citations identified through electronic database searching. All screening decisions were made and tracked in a Distiller SR database (Evidence Partners Inc., Manotick, ON, Canada).

Data Extraction

The research team created data abstraction forms and evidence table templates for each KQ. Based on clinical and methodological expertise, a pair of investigators was assigned to abstract data from each eligible article. One investigator abstracted the data, and the second reviewed the

completed abstraction form alongside the original article to check for accuracy and completeness. Disagreements were resolved by consensus, or by obtaining a third reviewer's opinion if consensus could not be reached.

We designed the data abstraction forms to collect the data required to evaluate the specified eligibility criteria for inclusion in this review, as well as demographic and other data needed for determining outcomes (screening test performance characteristics, as well as intermediate, final, and adverse events outcomes). We paid particular attention to describing the details of the screening intervention that may be related to outcomes, including setting, provider, timing, and frequency of screening; patient characteristics (e.g., age, parity); and study design (e.g., randomized controlled trial [RCT] versus observational). In addition, we described comparators carefully, as intervention and assessment standards may have changed during the study period. Harms outcomes were framed to help identify adverse events (e.g., stigmatization, decreased quality of life). Data necessary for assessing quality and applicability were also abstracted. Before the data abstraction form templates were used, they were pilot-tested with a sample of included articles and revised as necessary.

Quality Assessment of Individual Studies

We assessed the methodological quality, or risk of bias, of individual studies using the assessment instruments detailed in the *Methods Guide*[5] and *Medical Test Guide*.[6] To assess quality for studies presenting information on patient-centered intermediate, final, and adverse effect outcomes, we used a strategy to (1) classify the study design, (2) apply predefined criteria for quality and critical appraisal, and (3) arrive at a summary judgment of the study's quality. We applied criteria for each study type derived from core elements described in the *Methods Guide*. Criteria of interest for all studies included similarity of groups at baseline, extent to which outcomes were described, blinding of subjects and providers, blinded assessment of the outcome(s), intention-to-treat analysis, differential loss to followup between the compared groups or overall high loss to followup, and conflicts of interest. Criteria specific to RCTs included methods of randomization and allocation concealment. For observational studies, additional elements such as methods for selection of participants, measurement of interventions/exposures, addressing any design-specific issues, and controlling confounding were considered. To indicate the summary judgment of the quality of individual studies, we used the overall ratings of good, fair, or poor based on the study's adherence to well-accepted standard methodologies.

For studies assessing screening test performance elements within KQs 1, 2, and 3, we used QUADAS-2 (QUality Assessment of Diagnostic Accuracy Studies-2[8]) to assess quality. QUADAS-2 describes risk of bias in four key domains: patient selection, index test(s), reference standard, and flow and timing. The questions in each domain are rated in terms of risk of bias and concerns regarding applicability, with associated signaling questions to help with these bias and applicability judgments. Summary judgments for these studies were assigned as high risk of bias, low risk of bias, or unclear.

Data Synthesis

We began our data synthesis by summarizing key features of the included studies for each KQ. To the degree that data were available, we abstracted information on study design; patient

characteristics; clinical settings; interventions; screening test performance; and intermediate, final, and adverse event outcomes.

We determined the feasibility of completing a quantitative synthesis (i.e., meta-analysis) based on the volume of relevant literature, conceptual homogeneity of the studies (both in terms of study population and outcomes), and completeness of the reporting of results. We considered random-effects meta-analyses for comparisons where at least three conceptually homogenous studies reported the same patient-centered intermediate, final, or adverse effect outcome. Test performance was summarized using sensitivity, specificity, and likelihood ratios. When three or more conceptually homogeneous test performance studies were available, we considered random-effects bivariate meta-analysis to compute summary estimates of performance.

We anticipated that intervention effects might be heterogeneous. We hypothesized that the methodological quality of individual studies, study type, characteristics of the screening population (e.g., age, parity), and characteristics of the screening intervention (e.g., setting, provider) would be associated with the intervention effects. Where there were sufficient studies, we performed subgroup analyses and/or meta-regression analyses to examine these hypotheses.

To estimate the balance of benefits and harms of different screening strategies, we also adapted an existing simulation model of pregnancy and neonatal outcomes.[9] The model simulates pregnancy from conception through delivery and can subsequently simulate both maternal and child outcomes. We used the estimated likelihood of specific outcomes as the model output. Specific benefits include estimates of treated depression, false negatives, and false positives. Despite sparse data for harms, we can readily estimate the number of false-positive screening test results, or total referrals for further evaluation, under different scenarios. This allows an approach which compares total tests or false-positive results as a measure of "cost" or "harm" with a measure of benefit, such as "cases of depression detected."

The values for sensitivity and specificity (along with CIs) were derived from the literature review. The model also incorporates variability in followup and appropriate treatment after a positive screening test result. We used probabilistic sensitivity analysis to assess overall uncertainty based on the available literature, and used a modified value-of-information approach to help prioritize future research needs.[10] Because the report found almost no evidence from which to derive estimates for longer term outcomes, we focused the analysis on estimating the number of detected cases of depression, false negative and false positive results under different scenarios of test performance, and prevalence of depression.

Strength of the Body of Evidence

We rated the strength of evidence for each KQ and outcome using the approach described in the *Methods Guide*[5,11] and *Medical Test Guide*.[6] In brief, the approach requires assessment of four domains: risk of bias, consistency, directness, and precision. Additional domains were used when appropriate, namely, strength of association (magnitude of effect) and publication bias. These domains were considered qualitatively, and a summary rating of "high," "moderate," or "low" strength of evidence was assigned after discussion by two reviewers. In some cases, high, moderate, or low ratings were impossible or imprudent to make; for example, when no evidence was available or when evidence on the outcome was too weak, sparse, or inconsistent to permit any conclusion to be drawn. In these situations, a grade of "insufficient" was assigned.

Applicability

We assessed applicability across our KQs using the method described in the *Methods Guide*[5,11] and *Medical Test Guide;*[6] In brief, this method uses the PICOTS format as a way to organize information relevant to applicability. Items of particular interest that may contribute to heterogeneity and impact applicability include setting (e.g. country, provider), comparator, spectrum of disease (e.g. screening population or pre-selected), patient income, race, ethnicity, parity, and partner support. We used checklists to guide the assessment of applicability. We used these data to evaluate the applicability to clinical practice, paying special attention to study eligibility criteria, demographic features of the enrolled population in comparison to the target population, characteristics of the intervention used in comparison with care models currently in use, and clinical relevance and timing of the outcome measures. We summarized issues of applicability qualitatively.

Results

We begin by describing the results of our literature searches and then provide a brief description of the included studies. The remainder of the section is organized by Key Question (KQ). Under each of the six KQs, we begin by listing the key points of the findings, followed by a brief description of included studies and a detailed synthesis of the evidence. We did not conduct any quantitative syntheses.

Searches of PubMed, Embase, PsycINFO, and CDSR yielded 4,163 citations, 1,127 of which were duplicate citations. Manual searching identified 150 additional citations, for a total of 3,186 citations to be screened. After applying inclusion/exclusion criteria at the title-and-abstract level, 1,183 full-text articles were retrieved and screened. Of these, 1,143 were excluded at the full-text screening stage, leaving 40 articles for data abstraction. These 40 articles described 36 unique studies. The relationship of studies to the review questions is as follows: 15 studies relevant to KQ 1, 14 studies relevant to KQ 2, 2 studies relevant to KQ 3, 4 studies relevant to KQ 4, 1 study relevant to KQ 5, and 5 studies relevant to KQ6 (some studies were relevant to more than one KQ).

KQ 1. Performance Characteristics of Screening Instruments

We identified 15 studies that met the inclusion criteria for KQ 1. All confirmed the diagnosis of depression using a validated clinical interview or diagnostic instrument in screen positives and all or a sample of screen negatives. Four studies were performed in the United States, five in Europe, three in the UK, and one each in Australia, Asia, and Canada. Seven were judged to have a high risk of biased results; the remainder were judged to be at low risk.

Because no more than two studies provided results for the same test at the same threshold, we did not perform meta-analyses. Below, we present and discuss the results of the studies for each screening test qualitatively, then present the results for the three studies where two or more screening tests were directly compared. Only one study was relevant to KQ 1b.

Ten studies provided data on the Edinburgh Postnatal Depression Scale (EPDS), four on the Postpartum Depression Screening Scale (PDSS), two on the Beck Depression Inventory (BDI), and one each on the Antenatal Risk Questionnaire, the 17- and 21-Item Hamilton Rating Scale for Depression (HRSD-17 and HRSD-21), the Leverton Questionnaire, and the Patient Health Questionnaire (PHQ).

Table A summarizes the results and strength of evidence for each of the nine screening tests reviewed. In general, sensitivity estimates increased as specificity decreased, and sensitivity estimates were less precise than specificity estimates. For the majority of studies and tests, sensitivity and specificity estimates were in the 80–90 percent range. Because of the heterogeneity among studies in terms of setting, population, and choice of screening threshold, we were unable to perform quantitative synthesis, and confidence intervals (CIs) between tests broadly overlapped.

Table A. Strength of evidence domains for test characteristics of screening tests for postpartum depression

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Antenatal Risk Questionnaire						
Sensitivity	1 (276)	High	NA	Direct	Imprecise	Low SOE 78.1% (65.0 to 88.7%)
Specificity	1 (276)	High	NA	Direct	Imprecise	Low SOE 47.1% (40.3 to 59.9%)
BDI						
Sensitivity	2 (1,475)	Medium	Consistent	Direct	Imprecise	Low SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
Specificity	2 (1,475)	Medium	Consistent	Direct	Precise	Low SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
EPDS						
Sensitivity	10 (3,847)	Medium	Consistent	Direct	Imprecise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
Specificity	10 (3,847)	Medium	Consistent	Direct	Precise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
HSRD-17						
Sensitivity	1 (1376)	High	NA	Direct	Imprecise	Low SOE 80–85% (Range of point estimates across thresholds)
Specificity	1 (1,376)	High	NA	Direct	Imprecise	Low SOE 80–85% (Range of point estimates across thresholds)
HSRD-21						
Sensitivity	1 (1,375)	High	NA	Direct	Imprecise	Low SOE 80–85% (Range of point estimates across thresholds)
Specificity	1 (1,375)	High	NA	Direct	Imprecise	Low SOE 75–80% (Range of point estimates across thresholds)
Leverton Questionnaire						
Sensitivity	1 (617)	Low	NA	Direct	Imprecise	Low SOE 95.2% (90.4 to 98.1%)
Specificity	1 (617)	Low	NA	Direct	Imprecise	Low SOE 91.3% (88.4 to 93.7%)
PDSS						
Sensitivity	4 (904)	Medium	Consistent	Direct	Imprecise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
Specificity	4 (904)	Medium	Consistent	Direct	Precise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
PHQ						
Sensitivity	1 (506)	Low	NA	Direct	Imprecise	Low SOE 75–89% (Range of point estimates at varying threshold—wide 95% CIs for point estimates at each threshold)
Specificity	1 (506)	Low	NA	Direct	Imprecise	Low SOE 83–91% (Range of point estimates at varying threshold)
Two-Question Screen						
Sensitivity	1	Low	NA	Direct	Imprecise	Low SOE 100% (93.3 to 100%)
Specificity	1	Low	NA	Direct	Imprecise	Low SOE 44.3% (39.5 to 48.6%)

Abbreviations: BDI=Beck Depression Inventory; CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression; NA=not applicable; PDSS=Postpartum Depression Screening Scale; PHQ=Patient Health Questionnaire

KQ 2. Effect of Individual Factors on Screening Performance

We identified 15 articles describing 14 unique studies that met the inclusion criteria for KQ 2. Three were from the United States, seven were from Europe, two were Asian, and there was one study each from the UK, Australia, and Israel. Five were judged to be of good quality, with the rest being fair. We did not identify any studies relevant to KQ 2b. One study judged to be at high risk of bias did not provide an estimate of the association between parity and postpartum depression, but did provide separate estimates of screening test sensitivity and specificity stratified by parity.

Because of the inconsistency in how specific risk factors were described in the studies, we were unable to perform quantitative synthesis of the results. Table B presents the results from the included studies, and, except where noted, represents the results from each study's reported best fit multivariate model.

Table B. Strength of evidence domains for associations with patient characteristics and risk of postpartum depression

Risk Factor	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Maternal demographics						
Age	2 (4,773)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Education	2 (4,773)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Income	2 (4,998)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Employment status (unemployed vs employed)	1 (363)	High	NA	Direct	Imprecise	Low SOE OR 2.8 (1.1. to 4.9)
Obstetric history						
Parity	2 (4,998)	Medium	Consistent	Direct	Imprecise	Insufficient
Preterm/low birthweight infant	2 (4,711)	Medium	Consistent	Direct	Precise	Low SOE for Increased risk of postpartum depression
Smoking	2 (4,773)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Alcohol use	1 (4,348)	Medium	NA	Direct	Imprecise	Insufficient
General medical history						
Poor health status/chronic illness	2 (4,773)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Obesity	1	Medium	NA	Direct	Imprecise	Insufficient
Psychiatric history						
History of perinatal depression	2 (640)	High	NA	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
History of depression	5 (1,440)	Medium	Consistent	Direct	Precise	Moderate SOE for Increased risk of postpartum depression
History of premenstrual dysphoric disorder	1	Medium	NA	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Any psychiatric diagnosis	2 (1,025)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Anxiety	1	Medium	NA	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Personality (vulnerable/ neuroticsm)	2 (777)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression

Risk Factor	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Relationship/social support						
Marital status (single/no relationship)	2 (4,773)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Poor relationship quality	5 (5,856)	Medium	Consistent	Direct	Imprecise	Moderate SOE for Increased risk of postpartum depression
Poor social support	4 (1,333)	Medium	Consistent	Direct	Imprecise	Moderate SOE for Increased risk of postpartum depression

Abbreviations: CI=confidence interval; NA=not applicable; OR=odds ratio; SOE=strength of evidence

Among potential maternal demographic risk factors, no statistically significant association was found between maternal age, education, income, or type of employment. One study did, however, find a significant association between maternal unemployment and postpartum depression (OR 2.8; 95% CI, 1.1 to 4.9), although the overall strength of evidence was considered low.

Having a preterm or very low birthweight baby were both significantly associated with postpartum depression. In another study, having a 2nd or 3rd trimester termination for severe fetal abnormalities was associated with an increased risk of depression 14 months after the event compared with women with healthy infants, but there was no comparison with women with children with severe abnormalities who did not terminate the pregnancy.

Among potential general medical history risk factors, fair/poor self-reported health status and a history of chronic illness outside of pregnancy both increased the risk of postpartum depression over two-fold.

Past history of depression or anxiety, including both postpartum and before pregnancy, were consistently associated with an increased risk of postpartum depression, with odds ratios well above 2.0. Two studies also found that certain personality traits (neuroticism, vulnerability, low organization) were risk factors for depression.

Finally, although studies used a variety of different scales to measure the effect of relationship quality and social support on risk of depression, and were conducted in a wide range of settings ranging from urban United States to Singapore, the qualitative results were consistent: postpartum depression was significantly more common among women in poorer quality relationships (or no relationship), and among women with poor social support.

KQ 3. Effect of Testing Variables (Timing, Frequency, Setting, Provider) on Screening Performance

Two studies met the inclusion criteria for timing. No studies were identified that met the inclusion criteria for setting or provider.

Of the two KQ 3a studies, one was U.S. federally funded. Women participating in this prospective investigation of maternal mental illness at a single academic center were enrolled prior to 28 weeks gestation and followed through 6 months postpartum. Participants completed the EPDS, BDI, HRSD-17, and HRSD-21 during six perinatal windows: preconception, first

trimester, second trimester, third trimester, early postpartum (0–6 weeks), and later postpartum (7–26 weeks). The diagnosis of depression was confirmed by the Mood Module of the Structured Clinical Interview for Depression (SCID).

The second study was funded by a combination of industry and philanthropic monies. This single-center prospective study in Dublin, Ireland, enrolled women during the immediate postpartum period to determine if the EPDS, administered prior to hospital discharge, was predictive of depression at 6 weeks postpartum. Nine hundred fifty-one enrolled women completed the EPDS at 3–5 days postpartum with planned followup at 6 weeks postpartum for repeat EPDS and diagnostic interview using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) if screen-positive.

Neither study identified a significant effect of timing on test characteristics (Table C).

Table C. Strength of evidence domains for the effect of varying timing, setting, and provider on screening for postpartum depression

Variable	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Timing						
Delivery to 8 weeks vs 8 weeks to 6 months	1 (534)	High	NA	Direct	Imprecise	Insufficient
Delivery vs 6 weeks	1 (625)	High	NA	Direct	Imprecise	Insufficient

Abbreviations: CI=confidence interval; NA=not applicable; SOE=strength of evidence

KQ 4. Comparative Benefits of Screening, and KQ 5. Comparative Harms of Screening

Four studies met our inclusion criteria and evaluated the comparative benefits of screening for postpartum depression. Three were RCTs, and one was a quasi-experimental study. In terms of quality, one each of the three RCTs was rated as poor, fair, and good. The quasi-experimental study was rated as poor in quality. The most common relevant outcome was change in a screening instrument depression score. Sample size ranged from 99 recruited at one site to 7,649 enrolled from 101 practices. Only one study was conducted in the United States; it included only women receiving public assistance and at risk for postpartum depression. One each of the remaining studies was conducted in the UK, Norway, and Hong Kong. One of these studies was the only one to provide any evidence of harms.

Table D summarizes the strength of evidence and findings. Only two studies directly compared screening with no screening. One fair quality RCT found improvement in EPDS scores at 6 months in women randomized to screening at 2 months post-delivery compared to women randomized to no screening, but no differences in other measures, including general maternal health or parental stress. The screened group was significantly more likely to have unscheduled doctor visits for their infants up to 6 months, but this difference was not significant in the 6-12 month period. A good quality RCT found improved overall mental health based on the SF-12 at 12 and 18 months in women randomized to screening, but no differences in other outcomes.

Table D. Strength of evidence domains for benefits and harms of screening for postpartum depression

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Benefits						
Depressive symptoms	3 (1171)	Medium	Inconsistent	Direct	Imprecise	Low SOE for Reduced number of symptoms with screening and intervention
Mental health score (SF-12)	1 (418)	Low	NA	Direct	Imprecise	Low SOE for Improved scores with screening and intervention
Parental stress	2 (853)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Harms						
Unscheduled doctor visits for infant	1 (430)	Medium	NA	Direct	Imprecise	Low SOE for Increased number of visits for infants of screened women

Abbreviations: CI=confidence interval; NA=not applicable; SF-12=Medical Outcomes Study 12-Item Short-Form Health Survey; SOE=strength of evidence

KQ 6. Factors Affecting the Likelihood of an Appropriate Action After a Positive Screening Result

Five studies met the inclusion criteria for KQ 6. Two were prospective cohort studies, one was a cross-sectional study, one was a pre-post intervention study, and one had a quasi-experimental design. The two cohort studies were rated as fair quality, the cross-sectional study was rated as good quality, while the pre-post intervention study and quasi-experimental study were rated as poor quality. Four studies were conducted in the United States and one in Germany. All five provided some measure of appropriate diagnosis and treatment of depression. Screening most commonly occurred in the first 8 weeks postpartum; four of the five studies used the EPDS as the screening tool.

The main finding of these studies was that followup rates for women with positive screening tests were low, ranging from 5–30 percent (Table E). Referral was significantly higher in women with abnormal screening test results during the delivery admission compared with 36 weeks gestation or 6 weeks postpartum.

Table E. Strength of evidence domains for the effect of timing, setting, and provider on rates of referral and treatment among women with a positive screening test for postpartum depression

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Timing						
Prenatal vs postpartum	2 (784)	Medium	Consistent	Direct	Imprecise	Low SOE for Higher rates of referral/diagnosis prenatally
Delivery vs postpartum	1 (293)	Low	NA	Direct	Imprecise	Low SOE for Higher rates of referral/diagnosis during delivery admission

Abbreviations: CI=confidence interval; NA=not applicable; SOE=strength of evidence

Discussion

Findings in Light of Other Studies

Our review focused on studies published subsequent to the 2005 AHRQ Evidence Report on perinatal depression.[2,3] Our findings were largely consistent with that report. Although there was some new evidence addressing a few of the research gaps identified in that report (including more studies in ethnically diverse U.S. populations, direct comparisons of different screening instruments within studies, and direct comparisons of outcomes in screened vs. unscreened women), the strength of the additional evidence did not allow any conclusions about the overall balance of benefits and harms.

Applicability

The effects of interventions, as determined in research studies, do not always translate well to usual practice, where patient characteristics, clinical training, diagnostic workup, and resources may differ importantly from study conditions. Thus, we assessed the applicability of the included studies.[12]

Many included studies recruited populations where large differences between demographics of study population and community patients existed. Overall, only about one-third of included studies were conducted in the United States; the majority were conducted in Europe or the UK (48%). Event rates for postpartum depression between countries differ significantly due to dissimilarities in social and cultural contexts (e.g., family structures, gender roles). Moreover, the health care system in the United States differs considerably from those in Europe and the UK, making it problematic to translate findings to the U.S. context. Many studies had highly selected samples due to high rates of non-response or attrition during the studies, which limits these findings to broader populations. The majority of studies were conducted in women in their late twenties to early thirties. Few studies were conducted with samples of older maternal age. Finally, the prevalence of major depression in studies estimating the sensitivity and specificity was substantially higher than U.S.-population-based point-prevalence estimates, suggesting that that the positive predictive value of any screening instrument in a low-risk population will be substantially lower than the estimates derived from validation studies.

The EPDS is the most widely known and used screening tool for postpartum depression: over two thirds of studies assessed postpartum depression with the EPDS. To the extent that the EPDS is considered "standard of care," findings from these studies would have reasonable applicability. However, these studies used a range of cutoffs to signal probable postpartum depression (range: 8–13), and descriptions of testing protocols were not specific enough to inform routine clinical care. Confidence intervals for sensitivity estimates for all screening tests were wide, and for the most part sensitivity and specificity estimates were qualitatively similar. In addition, some studies administered the screening test in the perinatal through discharge period in a hospital setting—the results from this setting may not be representative of the results for screening in outpatient settings.

There were few direct comparisons between screening instruments, and the studies that did directly compare instruments did not identify substantial differences. There were only a few studies that directly compared screening with any instrument with no screening, and, although they suggest an improvement in depressive symptoms, there is limited data on other maternal or infant health outcomes. Lastly, there is limited information on paternal outcomes.

Implications for Clinical and Policy Decisionmaking

The 2005 AHRQ report concluded that there was a lack of evidence on the overall effectiveness of screening for depression in pregnancy or the postpartum period, lack of consensus on the appropriate target for screening (major depression alone vs. major and minor depression), and, if screening is performed, uncertainty about which instrument to use. These uncertainties are reflected in the recommendations by various stakeholder organizations discussed in the Introduction, and the evidence reviewed for this report does little to resolve those uncertainties: we found some evidence that screening improves some maternal outcomes compared with no screening, but the overall effect of this improvement on longer term maternal and infant outcomes is unclear.

The evidence is insufficient to guide recommendations about whether or not to screen for postpartum depression, and, if so, the optimal strategy to use. The relatively low prevalence of depression suggests that serial testing with a highly sensitive test followed by a highly specific test might be particularly efficient. Using a stochastic model, we simulated the trade-offs between false positive and true positive results, using sensitivity and specificity estimates for the two-question screen, the EPDS, and the PPDS. Serial testing with the two-question screen, followed by the use of the EPDS or PPDS in women screening positive to one of the two questions, almost always resulted in a lower false positive/true positive ratio compared with single screening with either the EPDS or PPDS alone, but this needs to be confirmed with empiric data. Modeling results suggest that better insight into what the specific benefits and harms of screening are, and how these should be weighted, may help guide choices about screening strategy.

Research Gaps

General Gaps

As noted above, one of the major limitations of the current evidence base is the wide disparity in methods and definitions used in studies relevant to screening for postpartum depression. This disparity limits the ability to synthesize the existing literature across disciplines;

in particular, it significantly limits the ability to perform meta-analyses. It would be extremely valuable for researchers in the field to reach consensus on a core set of measures that would be reported consistently across all relevant studies. For studies of interventions, common outcomes measures are the highest priority. For observational studies, or other study designs where there is a need to adjust for potential confounding, common measures for both outcomes and confounders are needed. In practice, this means not only agreement on *which* variables to collect, but *how* to measure and report them. For example, parity is frequently reported as a mean and standard deviation, which is not only clinically meaningless (since non-integer values of number of deliveries have no interpretation), but does not reflect the underlying distribution.

For many of the recommendations below, use of formal simulation and decision models may prove useful. As illustrated above, even a simple model can be useful for illustrating trade-offs, and can highlight the relationship between uncertainty about the relative likelihood of adverse outcomes compared to favorable outcomes, the acceptable harm/benefit trade-off, and the extent to which further research will help clarify the optimal decision or recommendation. This approach can be done using both specific clinical outcomes, or it can explicitly incorporate costs; in the latter case, this value-of-information analysis can help inform research prioritization and research budgeting.[10,13] Further development of the model outlined in this report could incorporate variations in strategies, such as timing of screening relative to delivery, repeated screening at varying intervals during pregnancy and the postpartum period, use of strategies to target high risk groups for screening, and strategies to enhance followup and treatment of women with positive screening results.

KQ 1

- Although greater precision for sensitivity estimates would be useful, there will always be greater uncertainty about sensitivity than specificity in a screening setting, since the number of subjects with the underlying condition will always be much smaller than the number of subjects without the condition. Given this limitation, it would ultimately be more efficient to perform studies large enough to address the question directly rather than multiple additional smaller studies, particularly if the smaller studies focus on a single instrument. We would suggest
    1. Achieving consensus on the appropriate trade-off between false positives and false negatives and using thresholds defined by these clinical criteria to determine optimal sensitivity and specificity for candidate screening instruments.
    2. Determination of other criteria for evaluating screening instruments (ease of administration, time associated with administration, costs, patient and provider acceptability, etc.). These criteria could be collected as part of the study. Alternatively, patient and provider acceptability could be measured using methods such as discrete choice experiments to assess the relative importance of different attributes of the screening test;[14] these data could then be used to inform the choice of which instruments to evaluate further.
    3. Define sample size for the study based on detecting clinically relevant differences in test performance and acceptability, with these differences being at least partially derived empirically in the first two steps.
    4. Direct comparison of candidate instruments, either by having the same subject use each instrument (randomized as to order of administration), or randomizing different subjects to different instruments. The trade-off here is between the

increased generalizability of having subjects take a single test versus overall sample size.

- The question of whether different instruments are better at identifying specific signs and symptoms is only important if there are effective interventions for those specific signs and symptoms. Clarity is needed on which signs and symptoms, and what potential interventions are available, in order to discuss potential research designs. One first step might be a systematic review focused on the individual signs and symptoms identified in the different screening instruments, with an emphasis on identifying effective interventions.

KQ 2

- Although we identified a number of consistent risk factors for postpartum depression, we did not identify any articles that used a multivariate predictive model to stratify patients by risk of developing the condition in order to screen more efficiently (similar to the Gail model used for identifying higher risk women for more aggressive screening protocols). The potential impact of such a model could be estimated by estimating the absolute risk of postpartum depression at different thresholds and using this information to estimate the number of false positives and false negatives resulting from screening only women identified as high risk. This could be compared to the estimated number of unwanted screening outcomes resulting from other strategies designed to minimize false positives, such as serial testing, using a simulation model. These data could, in turn, be used to estimate the size, costs, and value-of-information of a comparative trial.

KQs 3–6

- Additional RCTs comparing screening with no screening are needed. Ideally, some of these studies could address issues relevant to differences in timing, setting, or provider, perhaps through factorial designs.
- Explicit definitions of harms and benefits are needed.
- Ideally, these studies should include a long-term followup component for both mothers and infants. Although this will substantially affect costs and timing of the studies, if the ultimate rationale for screening involves both maternal and child outcomes, then a more explicit demonstration of the benefits in terms of these longer term outcomes is needed.
- There was insufficient evidence to address the effect of timing, setting, or provider on test characteristics, while there was low strength evidence that timing might affect likelihood of receiving appropriate diagnostic and therapeutic services. It seems plausible that differences in outcome relevant to timing, setting, or provider are more related to aspects of the screening/referral/diagnosis process other than the test characteristics of the specific screening instrument used in the study.

Conclusions

Although currently available screening instruments are reasonably sensitive and specific in detecting postpartum depression, there is insufficient evidence to draw any conclusions about the net balance of benefits and harms of screening for postpartum depression, or about whether specific tools or strategies would result in a more favorable balance.

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR). Arlington, VA: American Psychiatric Association; 2000.

2. Gaynes BN, Gavin N, Meltzer-Brody S, et al. Perinatal Depression: Prevalence, Screening Accuracy, and Screening Outcomes. Evidence Report/Technology Assessment No. 119 (Prepared by RTI–University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016). Rockville, MD: Agency for Healthcare Research and Quality; February 2005. AHRQ Publication No. 05-E006-2.

3. Gaynes BN, Gavin N, Meltzer-Brody S, et al. Perinatal depression: prevalence, screening accuracy, and screening outcomes. Evid Rep Technol Assess (Summ). 2005(119):1-8. PMID: 15760246.

4. Ng RC, Hirata CK, Yeung W, et al. Pharmacologic treatment for postpartum depression: a systematic review. Pharmacotherapy. 2010;30(9):928-41. PMID: 20795848.

5. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed January 3, 2012.

6. Agency for Healthcare Research and Quality. Methods Guide for Medical Test Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=558. Accessed January 3, 2012.

7. Anonymous. Evidence-based Practice Center Systematic Review Protocol. Project Title: Efficacy and Safety of Screening for Postpartum Depression. March 9, 2012. Available at: www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=997. Accessed June 21, 2012.

8. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36. PMID: 22007046.

9. Myers ER, Misurski DA, Swamy GK. Influence of timing of seasonal influenza vaccination on effectiveness and cost-effectiveness in pregnancy. Am J Obstet Gynecol. 2011;204(6 Suppl 1):S128-40. PMID: 21640230.

10. Myers E, Sanders GD, Ravi D, et al. Evaluating the Potential Use of Modeling and Value-of-Information Analysis for Future Research Prioritization Within the Evidence-based Practice Center Program. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-2007-10066-I.) AHRQ Publication No. 11-EHC030-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed January 3, 2012.

11. Owens DK, Lohr KN, Atkins D, et al. AHRQ Series Paper 5: Grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the Effective Health-Care Program. J Clin Epidemiol. 2010;63(5):513-23. PMID: 19595577.

12. Atkins D, Chang SM, Gartlehner G, et al. Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011;64(11):1198-207. PMID: 21463926.

13. Myers E, McBroom AJ, Shen L, et al. Value-of-Information Analysis for Patient-Centered Outcomes Research Prioritization. Prepared by the Duke Evidence-based Practice Center for the Patient-Centered Outcomes Research Institute. March 9, 2012. Available at: www.pcori.org/assets/Value-of-Information-

Analysis-for-Patient-Centered-Outcomes-Research-Prioritization.pdf. Accessed July 3, 2012.

14. Wordsworth S, Ryan M, Skatun D, et al. Women's preferences for cervical cancer screening: a study using a discrete choice experiment. Int J Technol Assess Health Care. 2006;22(3):344-50. PMID: 16984063.

Internet Citation

[To be provided in the final report.]

Introduction

Background

Postpartum Depression

Depression is a potentially life-threatening condition with a substantial impact on quality of life. The impact of depression in postpartum women is at least as great as that for depression in other populations. Postpartum depression is defined in the *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision* (hereafter, DSM-IV-TR) as a major depressive disorder according to the diagnostic criteria listed in Table 1, with a secondary criterion of onset of symptoms within 4 weeks of delivery.[1] A new set of diagnostic criteria for psychiatric illness, the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (DSM-V), is currently scheduled for release in May 2013; preliminary discussions suggest that the overall diagnostic framework for postpartum depression (i.e., major depression with a specification of postpartum onset) will remain unchanged, although the window for diagnosis may be extended to 6 months after delivery,[2] and "minor" will be replaced by "subthreshold."

Other diagnostic standards allow the definition of onset to extend beyond 4 weeks and up to 12 months after delivery and/or add a "minor depression" subcategory (2 to 4 of the symptoms listed in Table 1). There is high-quality evidence for effective treatment of patients who meet criteria for major depression in other settings; evidence is inconsistent for postpartum depression.[3-5]

Table 1. DSM-IV-TR diagnostic criteria for major depressive disorder

Criterion	Description
A.	Five (or more) of the symptoms below have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either 1) depressed mood or 2) loss of interest or pleasure. (Note: Do not include symptoms that are clearly due to a general medical condition, or mood-incongruent delusions or hallucinations.) • Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful) • Markedly diminished interest in pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others) • Significant weight loss when not dieting or weight gain (e.g., change of more than 5% body weight in a month), or decrease or increase in appetite nearly every day • Insomnia or hypersomnia nearly every day • Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down) • Fatigue or loss of energy nearly every day • Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick) • Diminished ability to think or concentrate, or indecisiveness, nearly every day (either subjective account or as observed by others) • Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide
B.	The symptoms do not meet the criteria for mixed episode (*DSM-IV-TR*, p.365)

Criterion	Description
C.	The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning
D.	The symptoms are not due to the direct physiological effects of a substance (e.g., a drug of abuse, medication) or a general condition (e.g., hypothyroidism)
E.	The symptoms are not better accounted for by bereavement, i.e., after the loss of a loved one, the symptoms persist for longer than 2 months or are characterized by marked functional impairment, morbid preoccupation with worthlessness, suicidal ideation, psychotic symptoms, or psychomotor retardation

Abbreviations: DSM-IV-TR=Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision

The most recent U.S.–based synthesis of the evidence, performed for the Agency for Healthcare Research and Quality (AHRQ) in 2005,[3,4] estimated that the prevalence of major depression alone during the first postpartum year is 1.0 to 5.9 percent, and that the prevalence of major and minor depression combined is 6.5 to 12.9 percent. Incidence estimates for the first 3 postpartum months were up to 6.5 percent for major depression alone and 14.5 percent for major and minor depression. At the time of the AHRQ review, consistent limitations in the literature included small sample size (precluding subgroup analyses) and lack of generalizability.

Although the risk of suicide in women may be lower during pregnancy and the postpartum period,[6] a review of maternal mortality in the United Kingdom during the 1990s found that suicide was the leading cause of maternal mortality, accounting for 29 percent of maternal deaths.[7,8] In addition, postpartum depression may increase the risk of infant mortality through neglect, abuse, or homicide.[9] Maternal depression clearly affects maternal-infant interactions and some measures of infant development.[10-13] Health care resource utilization is greater for women with postpartum depression than for postpartum women who are not depressed;[14] data on resource use for their infants are inconsistent.[15,16] Outcomes in the studies included in the two most recent systematic reviews were primarily scores on measures of depression, which are often used as endpoints in clinical trials of depression therapy; other clinical outcomes, such as measures of infant health, were not included.[3,4,17]

Screening for Postpartum Depression

Given the potential impact of postpartum depression on maternal and infant health, there has been considerable interest in strategies aimed at identifying women who are at risk for postpartum depression or who have postpartum depression, with the ultimate goal being the application of effective preventive or therapeutic interventions. Key components of any particular screening strategy for postpartum depression include (1) which screening test or instrument to use, (2) when to screen, (3) who should screen, and (4) how to use the results of the screening test. However, there is considerable uncertainty about all of these components, as seen in existing recommendations.

Potential Benefits of Screening

All major organizations providing care to pregnant and postpartum women and infants recognize the risk of postpartum depression and the potential benefit of screening, but the strength of recommendations is variable. For example, none of the U.S.–based organizations recommend use of a specific instrument (Table 2). Factors limiting the strength of recommendations include the lack of sufficient data on the most appropriate screening instrument, the optimal time(s) for screening,[18] issues concerning reimbursement and the scope

of practice,[10,18] and the need for adequate systems for ensuring appropriate care for women identified through screening.[10,11,19]

Table 2. Guidelines/recommendations for screening for postpartum depression

Organization	Statement	Date
U.S. Preventive Services Task Force[19]	No specific recommendations for postpartum depression. Grade B recommendation *for* screening "when staff-assisted depression care supports are in place to assure accurate diagnosis, effective treatment, and follow-up;" Grade C recommendation *against* screening when such supports are not in place.	December 2009
American College of Obstetricians and Gynecologists, Committee on Obstetric Practice[18]	At this time there is insufficient evidence to support a firm recommendation for universal antepartum or postpartum screening. There are also insufficient data to recommend how often screening should be done. However, screening for depression has the potential to benefit a woman and her family and should be strongly considered. Medical practices should have a referral process for identified cases. Women with current depression or a history of major depression warrant particularly close monitoring and evaluation.	February 2010
American Academy of Pediatrics, Committee on Psychosocial Aspects of Child and Family Health[20]	Screening can be integrated, as recommended by *Bright Futures* and the American Academy of Pediatrics Mental Health Task Force, into the well-child care schedule and included in the prenatal visit. This screening has proven successful in practice in several initiatives and locations and is a best practice for primary care pediatricians caring for infants and their families. Intervention and referral are optimized by collaborative relationships with community resources and/or by colocated/integrated primary care and mental health practices.	November 2010
American Academy of Family Physicians[21]	No specific recommendations for postpartum depression; general recommendations for screening follow those of the U.S. Preventive Services Task Force.[19]	October 2010
American College of Nurse Midwives[22]	The American College of Nurse Midwives supports universal screening, treatment, and/or referral for depression in women as a part of routine primary health care.	December 2003
United Kingdom National Institute for Health and Clinical Excellence[23]	At a woman's first contact with a primary care provider, at her booking visit, and postnatally (usually at 4 to 6 weeks and 3 to 4 months), healthcare professionals (including midwives, obstetricians, health visitors, and general practitioners) should ask two questions to identify possible depression: • During the past month, have you often been bothered by feeling down, depressed, or hopeless? • During the past month, have you often been bothered by having little interest or pleasure in doing things? A third question should be considered if the woman answers "yes" to either of the initial questions: • Is this something you feel you need or want help with? Health care professionals may consider the use of self-report measures such as the Edinburgh Postnatal Depression Scale (EPDS), the Hospital Anxiety and Depression Scale (HADS), or the Patient Health Questionnaire-9 (PHQ-9) as part of a subsequent assessment or for the routine monitoring of outcomes.	April 2007

Abbreviations: EPDS=Edinburgh Postnatal Depression Scale; HADS=Hospital Anxiety and Depression Scale; PHQ-9=Patient Health Questionnaire-9

Despite this uncertainty, efforts have been made at the state level to require offering screening for postpartum depression, although the experience to date has not demonstrated substantial benefit.[24]

Potential Harms of Screening

In their 2009 recommendations on screening for depression in adults, the U.S. Preventive Services Task Force identified "false-positive results, the inconvenience of additional diagnostic workup, the costs and adverse effects of treatment of patients who are incorrectly identified as being depressed, and potential adverse effects of labeling" as potential harms but found no evidence for any of these harms.[19] Whether any of these harms is more likely when screening for postpartum depression is unclear. However, it is possible that pregnant and postpartum women may be at increased risk of harm from screening, given that many of the signs and symptoms included in the diagnostic criteria for depression (Table 1) are common and normal responses to pregnancy, childbirth, and caring for infants. Furthermore, many studies of postpartum depression include "minor depression" as a diagnostic category. Previous reviews have concluded that there is a lack of evidence that treatment of symptoms not meeting criteria for major depression improves outcomes.[3,4,25] If a diagnosis of "minor depression" does not lead to effective treatment, then patients are exposed to the potential side effects of therapy (particularly medical therapy) in addition to being labeled as depressed without a concomitant improvement in health for themselves or their child. Finally, when comparing different strategies for screening women, differences in both false-positive and false-negative results are important, especially for women who might have been helped by earlier identification of depression through screening.

Accuracy of Screening Instruments

In evaluating strategies involving screening for postpartum depression, patients, providers, and policymakers must consider the tradeoffs between the likely benefits and harms of screening. Although direct evidence from appropriately designed trials is ideal, such data are often lacking (and are lacking for screening for postpartum depression). In such cases, inferences must be drawn from data on how well the screening test or strategy distinguishes between patients who truly have the condition of interest and those who do not, which is usually reported as the strategy's sensitivity (the likelihood that people with the condition will have a positive test) and specificity (the likelihood that people without the condition will have a negative test). The sensitivity and specificity of a test are characteristics that are independent of the population being tested. Higher sensitivity means fewer people with the condition are missed, while higher specificity means fewer people without the condition will be falsely identified; importantly, sensitivity and specificity are indirectly correlated—increasing sensitivity decreases specificity and vice versa. One advantage of sensitivity and specificity is that, because they are characteristics of the tests themselves, sensitivity and specificity estimates of a given test can be compared and pooled across different studies.

Sensitivity and specificity are not, however, directly useful clinically; the more relevant test characteristics are positive predictive value (PPV; the likelihood that a person with a positive test has the condition of interest) and negative predictive value (NPV; the likelihood that a person with a negative test does not have the condition of interest). These characteristics are functions of test sensitivity and specificity *and* the underlying likelihood of the condition of interest (prevalence). Because of this dependence on prevalence, the PPV and NPV of a specific test can vary across studies, depending on the population. The PPV and NPV of a test or strategy can be

directly estimated from a study in a specific population or can be indirectly estimated from given estimates of the test sensitivity and specificity and the population prevalence. A test with a certain sensitivity and specificity might have quite different PPV and NPV when used in different settings or at different times. Greater certainty about how PPV and NPV vary across populations, settings, and timing would help in developing specific recommendations about when, whom, and how often to screen.

One of the consistent uncertainties identified in current postpartum screening recommendations is how well currently available tests and strategies for identifying women with, or at risk for, postpartum depression perform. For example, the committee opinion on screening for depression during and after pregnancy developed by the American Congress of Obstetricians and Gynecologists[18] lists seven different tests—the Edinburgh Postnatal Depression Scale (EPDS), the Postpartum Depression Screening Scale (PDSS), the Patient Health Questionnaire-9 (PHQ-9), the Beck Depression Inventory (BDI), the Beck Depression Inventory II (BDI-II), the Center for Epidemiologic Studies Depression Scale (CES-D), and the Zung Self-Rating Depression Scale (Zung SDS)—with wide ranges for the reported sensitivity and specificity, but it does not provide specific guidance on which test might be most appropriate in a particular setting.

Another issue is that sensitivity and specificity may also vary based on the definition of "disease." For example, the 2005 AHRQ Evidence Report on postpartum depression[3,4] found that the sensitivity of all instruments reviewed was greater for "mild" depression when compared with "major" depression." As noted above, if treatment of mild depression does not lead to improved outcomes, then this greater "sensitivity" does not translate into a better test.

Clinical and Socioeconomic Factors Affecting Risk for Postpartum Depression

Consistent risk factors for postpartum depression identified in the literature include a history of depression before pregnancy, depression or anxiety during pregnancy, experiencing stressful life events during pregnancy or the early postpartum period, and low levels of social support; maternal age, income, and parity may also affect risk.[26-30] Because the outcomes of screening for any condition are dependent on the likelihood of that condition at the time of screening, selective use of specific tools to screen women at higher risk for postpartum depression when one or more risk factors are present may be a viable strategy.

Other Factors Affecting Screening Performance

Timing

Many of the signs and symptoms that make up the diagnostic criteria for depression are also common physiological or emotional responses to pregnancy and caring for an infant, and their prevalence can vary depending on when the measurement is performed. The presence of similar signs/symptoms in women who have and do not have depression could affect the specificity, and thus the false-positive rate, of a given screening test. In addition, testing during the prenatal period is seeking either to identify current depression (which by definition would not be postpartum depression), or to identify women *at risk* for postpartum depression; the performance of a test designed to identify patients at higher risk before they develop a condition is often quite different than the performance of a test designed to detect the condition itself.

Setting

Setting is inevitably related to timing; however, setting may have other effects on test performance. For example, the willingness of a woman to admit to symptoms of depression might vary depending on the setting—that is, her comfort level and familiarity with a provider or her concerns about being judged as a parent. Setting may also play a crucial role in determining whether women with a positive screening test result receive appropriate diagnostic and treatment services.

Provider

As with setting, the provider and the nature of his/her relationship with the patient may affect the willingness of the patient to admit to symptoms of depression. The provider's ability to appropriately administer a given screening tool may be affected by his/her training or the nature of his/her usual practice. Finally, as with setting, even if the sensitivity/specificity/predictive values of the test are unchanged, the ability of the provider to provide appropriate diagnosis and treatment to a patient with a positive test may vary based on available resources, skill and training of provider, or the context of visit.

Effective Management of Positive Screening Tests

Screening is often focused during pregnancy or the first 3 postpartum months in settings where care is provided to pregnant or postpartum women by providers such as obstetricians, family practitioners, or nurse-midwives. All of the existing recommendations for screening emphasize the need for systems or procedures to ensure that women identified as being at risk for postpartum depression receive appropriate diagnostic services, and, if a diagnosis of depression is confirmed, appropriate treatment (Table 2). Because the risk of postpartum depression extends throughout the first 12 months after delivery, maternal depression may affect outcomes for the infant, and settings where care is provided to the infant provide an opportunity for postpartum depression screening. Clinicians who provide care for infants have proposed the possibility of including screening for maternal depression as part of routine infant care,[11,20] but issues regarding scope of practice, legal liability, and appropriate referral remain challenges.[10]

Scope and Key Questions

Scope of the Review

This Comparative Effectiveness Review (CER) was funded by AHRQ and is designed to evaluate the comparative diagnostic accuracy, benefits, and harms of available screening instruments for postpartum depression. Further details are provided under "Key Questions" and "Analytic Framework," below, and in the section on "Inclusion and Exclusion Criteria" in the Methods chapter. For each Key Question (KQ), we further consider whether the diagnostic accuracy, benefits, and harms of the screening instruments evaluated differs among specific patient subgroups of interest, including subgroups defined by age, race, parity, history of mood disorders, history of intimate partner violence, perinatal outcomes, or cultural factors, or by the timing, setting, or provider.

Despite recognition that (a) postpartum depression is common, (b) it may have serious effects on both mothers and infants, and (c) screening instruments are available, uncertainty about whether, when, and how to screen for postpartum depression remains, as seen in the various recommendations summarized in Table 2. Sources for this uncertainty include:

- Imprecision in the published sensitivity and specificity estimates for the various instruments at the time the recommendations were drafted. Incorporating additional data published subsequently should add greater precision to these estimates by increasing the overall sample size and may make any differences between specific tests more apparent.
- Uncertainty about the ability of screening strategies to consistently identify the women most likely to benefit from available treatments and followup. For example, in populations at very low risk for postpartum depression, lower specificity would result in a low negative predictive value and could result in a high absolute number of women referred for additional diagnostic evaluation.
- Lack of direct evidence of benefits from screening. For screening to be of benefit, the test has to be able to accurately distinguish between those likely to benefit from further evaluation and treatment and those at low risk for the condition of interest; women identified as being at higher risk of the condition have to be able to receive appropriate diagnostic services; and, for those definitively identified with the condition, effective treatment needs to be available. Our review will focus on the first two aspects of screening benefits. If we assume that women identified through screening whose symptoms meet the diagnostic criteria for depression are given effective treatments, then a study that randomized women to no screening versus screening, or to screening with two different instruments, would address the question of screening benefit, especially if the treatments were standardized. Addressing the question of which treatments are most effective would require a different design.
- Issues related to management of women with a positive screening result. Although all recommendations related to screening commented on the need for appropriate systems or mechanisms for managing women with a positive screening test, there is no mention of the possible harms, such as anxiety created by a positive screening test result or the potential stigma associated with a diagnosis of depression.

By summarizing data that support improved screening of postpartum depression, we hope to enhance patient-centered outcomes for women, their partners, and children -- and to reduce health care utilization and costs. This project will benefit patients, providers, payers, and policymakers. Patients will benefit from more robust data on the comparative diagnostic accuracy, benefits, and harms of available screening instruments for postpartum depression. Providers will benefit by gaining a better understanding of which tools, timing, and setting are most effective for screening women for postpartum depression – and which patients benefit most from such screening. Policymakers will be able to design and implement programs to make better use of scarce health care resources while improving the health status of women at risk for postpartum depression.

Key Questions

With input from our Key Informants, we constructed Key Questions (KQs) using the general approach of specifying the populations, interventions, comparators, outcomes, timeframes, and settings of interest (PICOTS; see the section on “Inclusion and Exclusion Criteria” in the Methods chapter for details).

The KQs considered in this CER are:

KQ 1: This question has two parts:

a. What are the sensitivity and specificity of currently available screening instruments for detecting postpartum depression, and how do these translate into the likelihood of false-negative and false-positive results in different populations and settings?
b. Are there clinically relevant differences in the ability of currently available screening instruments to correctly identify specific signs or symptoms of depression (e.g., suicidal ideation)?

KQ 2: This question has two parts:

a. Are there individual factors (age, race, parity [number of live births], history of mood disorders, history of intimate partner violence, perinatal outcomes, cultural factors) that affect the baseline risk of postpartum depression and, therefore, the subsequent positive and negative predictive values of screening instruments?
b. Are there validated predictive models or algorithms based on such factors that would improve the performance of screening instruments?

KQ 3: Are the performance characteristics (sensitivity, specificity, predictive values) of screening instruments affected by:

a. Timing (prenatal, peripartum, or at various times in the first postpartum year) and frequency of screening?
b. Setting (prenatal visit, hospital/birthing center/home, postpartum maternal visit, or well-child visit)?
c. Provider (obstetrician, midwife, pediatrician, family practitioner, other health provider)?

KQ 4: What are the comparative benefits of screening for postpartum depression when compared to no screening, or between different screening strategies (based on choice of screening instrument, timing, setting, etc.)?

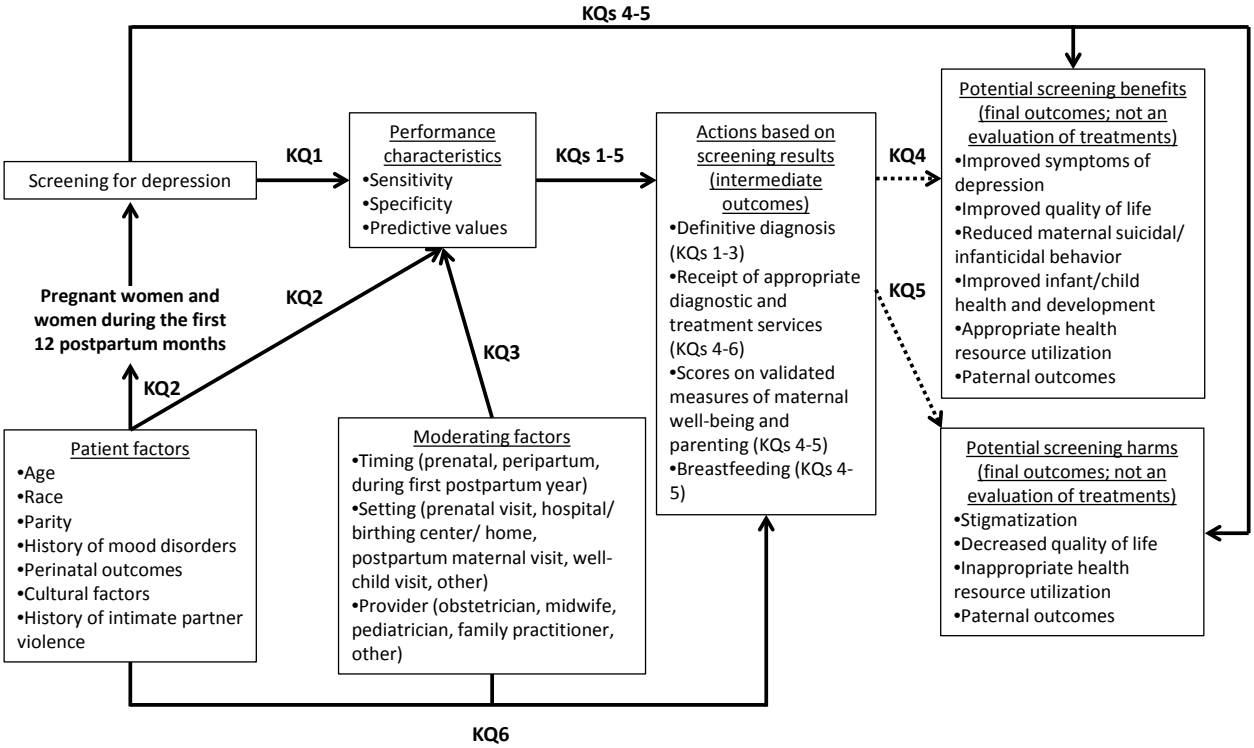
KQ 5: What are the comparative harms of screening for postpartum depression when compared to no screening, or between different screening strategies (based on choice of screening instrument, timing, setting, etc.)?

KQ 6: Is the likelihood of an appropriate action (referral, diagnosis, treatment, etc.) after a positive screening result affected by timing, setting, patient characteristics, or other factors?

Analytic Framework

Figure 1 shows the analytic framework for this project.

Figure 1. Analytic framework



Abbreviation: KQ=Key Question

This figure depicts the KQs within the context of the PICOTS described in the Methods section of this report. In general, the figure shows that the population of interest is pregnant women and women during the first 12 months postpartum. KQ 1 focuses on the sensitivity and specificity of currently available screening instruments for detecting postpartum depression. KQ 2 considers whether there are any individual factors (age, race, parity [number of live births], history of mood disorders, perinatal outcomes, cultural factors, and history of intimate partner violence) that affect the baseline risk of postpartum depression and therefore the subsequent positive and negative predictive values of screening instruments. KQ 3 considers whether the performance characteristics (sensitivity, specificity, and predictive values) of screening instruments are affected by the timing (prenatal, peripartum, or at various times in the first postpartum year), setting of administration (prenatal visit, hospital/birthing center/home, postpartum maternal visit, well-child visit, or other setting), or provider (obstetrician, midwife, pediatrician, family practitioner, or other health care provider). The outcome for KQs 1–3 is a definitive diagnosis of depression based on *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision* (DSM-IV-TR) criteria using a validated instrument. KQ 4 considers the potential benefits of screening for postpartum depression, including improved symptoms of depression, improved quality of life, reduced maternal suicidal or infanticidal behavior, improved infant/child health and development outcomes, and appropriate health resource utilization. KQ 5 considers possible harms associated with screening, including stigmatization, decreased quality of life, and inappropriate health resource utilization. Both KQ 4 and KQ 5 consider intermediate outcomes such as receipt of appropriate diagnostic and treatment services for symptoms of depression, scores on validated measures of maternal well-being and parenting, and breastfeeding. Paternal outcomes, including scores on validated mental health

instruments, health-related quality of life, and health system resource utilization, are also considered in both KQ 4 and KQ 5. KQ 6 asks whether the likelihood of an appropriate action (defined as receipt of appropriate diagnostic and treatment services for symptoms of depression) after a positive screening result is affected by the same timing, setting, and patient characteristic variables considered in KQs 2 and 3.

Methods

The methods for this Comparative Effectiveness Review (CER) follow those suggested in the Agency for Healthcare Research and Quality (AHRQ) *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (hereafter referred to as the *Methods Guide*)[31] and *Methods Guide for Medical Test Reviews* (hereafter referred to as the *Medical Test Guide*).[32] The main sections in this chapter reflect the elements of the protocol established for the CER; certain methods map to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.[33]

Topic Refinement and Review Protocol

During the topic refinement stage, we solicited input from Key Informants representing medical professional societies/clinicians in the areas of mental health, obstetrics and gynecology, women's health, pregnancy and perinatal epidemiology, psychiatry, maternal and fetal medicine, pediatrics, and primary care; patients, scientific experts; and payers, to help define the Key Questions (KQs). The KQs were then posted for public comment for 4 weeks from November 8 to December 6, 2011, and the comments received were considered in the development of the research protocol. We next convened a Technical Expert Panel (TEP) comprising clinical, content, and methodological experts to provide input in defining populations, interventions, comparisons, and outcomes, and in identifying particular studies or databases to search. The Key Informants and members of the TEP were required to disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts. Any potential conflicts of interest were balanced or mitigated. Neither Key Informants nor members of the TEP performed analysis of any kind, nor did any of them contribute to the writing of this report. Members of the TEP were invited to provide feedback on an initial draft of the review protocol which was then refined based on their input, reviewed by AHRQ, and posted for public access at the AHRQ Effective Health Care Website.[34]

Literature Search Strategy

Search Strategy

To identify relevant published literature, we searched PubMed®, Embase®, PsycINFO®, and the Cochrane Database of Systematic Reviews (CDSR), limiting the search to studies published from January 1, 2004, to January 25, 2012 (subsequent to the March 2004 search end date of the 2005 AHRQ evidence report on postpartum depression).[3,4] Where possible, we used existing validated search filters (such as the Clinical Queries Filters in PubMed). An experienced search librarian guided all searches. Exact search strings are included in Appendix A. We supplemented the electronic searches with a manual search of citations from a set of key primary and systematic review articles.[4,11,17,28,29,35-63] All citations were imported into an electronic database (EndNote® X4; Thomson Reuters, Philadelphia, PA).

As a mechanism to ascertain publication bias, we searched ClinicalTrials.gov to identify completed but unpublished studies. While the draft report is under peer review, we will update the literature search and include any eligible studies identified either during that search or through peer or public reviews in the final report.

We used several approaches to identify relevant grey literature; these included requests to publishers of proprietary depression screening tools (from among those listed in Table 3) for scientific information packets and searches of trial registries and conference abstracts for relevant articles from completed studies. Grey literature databases included ClinicalTrials.gov; the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) search portal; and ProQuest COS Conference Papers Index. Search terms used for these sources are provided in Appendix A.

Inclusion and Exclusion Criteria

The PICOTS (Populations, Interventions, Comparators, Outcomes, Timings, and Settings of interest) criteria used to screen articles for inclusion/exclusion at both the title-and-abstract and full-text screening stages are detailed in Table 3.

Table 3. Inclusion and exclusion criteria

Study Characteristic	Inclusion Criteria	Exclusion Criteria
Populations	• Pregnant women and women up to 12 months postpartum • Subgroups of potential interest include: o Race/ethnicity o Income o Parity o Cultural norms o History of mood disorders o Perinatal outcomes o History of intimate partner violence • Fathers or domestic partners	• Women currently undergoing treatment for depression • Studies where the primary objective is to detect depression during pregnancy (rather than to identify risk factors for postpartum depression) • Studies exclusively addressing bipolar disorder, a primary psychotic disorder, or maternity blues; or studies that include these populations and do not report results for subjects not fitting these subgroups separately

Study Characteristic	Inclusion Criteria	Exclusion Criteria
Interventions	• Screening using a validated screening instrument for depression, including, but not necessarily limited to: o Bromley Postnatal Depression Scale (BPDS) o Edinburgh Postnatal Depression Scale (EPDS) o Postpartum Depression Screening Scale (PDSS) o Leverton Questionnaire (LQ) o Center for Epidemiologic Studies Depression Scale (CES-D) o Hospital Anxiety and Depression Scale (HADS) o Patient Health Questionnaire-9 (PHQ-9) o Beck Depression Inventory (BDI-IA, BDI-II) o Zung Self-Rating Depression Scale (Zung SDS) o Hamilton Rating Scale for Depression (HRSD) o Postpartum Depression Predictors Inventory-Revised (PDPI-R) o General Health Questionnaire (GHQ-D) o Montgomery Asberg Depression Rating Scale (MADRS) o Generalized Contentment Scale o Patient Health Questionnaire-2 (PHQ-2) o Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MD PHQ)	• Validation studies, or screening conducted using a non-validated instrument
Comparators	• No formal protocol for screening, screening with another validated instrument, or screening with the same instrument under different conditions (e.g., different settings or different timing)	• Comparison to screening with a non-validated instrument
Outcomes	• Performance characteristics (KQs 1–3): o Sensitivity o Specificity o Predictive values • Intermediate outcomes o KQs 1–3: ▪ Diagnosis of depression based on the DSM-IV-TR criteria using a validated instrument o KQs 4 and 5: ▪ Receipt of appropriate diagnostic and treatment services for symptoms of depression ▪ Scores on validated measures of maternal well-being and parenting ▪ Breastfeeding o KQ 6: ▪ Receipt of appropriate diagnostic and treatment services for symptoms of depression • Final outcomes (KQ 4): o Scores on validated diagnostic instruments for depression o Health-related quality of life, based on validated measures o Maternal suicidal/ infanticidal behaviors	• Outcomes measured pre-delivery • Only outcome of interest reported is depression; outcome was measured with the screening instrument only and not confirmed with a reference standard • Only outcome of interest reported is sensitivity/specificity, and insufficient data provided to construct a 2-by-2 table • Article provides information only about the association between postpartum depression and other outcomes without linking screening to those outcomes

Study Characteristic	Inclusion Criteria	Exclusion Criteria
	o Scores on validated instruments of infant health and development o Maternal health system resource utilization, including number of visits and estimates of total and attributable costs o Infant health system resource utilization, including number of visits and estimates of total and attributable costs o Paternal outcomes, including scores on validated mental health instruments, health-related quality of life, and health system resource utilization (measured as described above for maternal outcomes) • Adverse effects (KQ 5): o Scores on validated measures of stigmatization o Health-related quality of life, based on validated measures o Maternal health system resource utilization, including number of visits and estimates of total and attributable costs o Infant health system resource utilization, including number of visits and estimates of total and attributable costs o Paternal outcomes, including scores on validated mental health instruments, health-related quality of life, and health system resource utilization (measured as described above for maternal outcomes)	
Timing	• Intervention o Prenatal period o Immediate postpartum period (up to 6 weeks after delivery) o Up to 12 months after delivery • Followup o Begins at delivery; timing of followup not limited	• Outcomes measured pre-delivery
Setting	• Any clinical provider setting, home • Study locations include at least one high-income economy as defined by the World Bank.[64] We restrict the study to economically developed countries—countries that have greater cultural and health care system similarities to the United States—to improve applicability of the study results to U.S. populations.	None
Study design	• Original data • RCTs, prospective and retrospective observational studies with comparator; for test characteristics, cross-sectional studies acceptable if includes patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard • RCTs: All sample sizes • Observational studies: sample size ≥100 subjects	• Editorials, nonsystematic reviews, letters, case series, case reports

Study Characteristic	Inclusion Criteria	Exclusion Criteria
Publications	• English-language only • Peer-reviewed articles • Relevant systematic review, meta-analysis, or methods article (to be used for background only)[a] • Published on or after January 1, 2004	• Non-English-language articles[b]

[a]Systematic reviews and meta-analyses were excluded from direct abstraction; those representing key sources were hand-searched as potential sources of additional citations to consider in the review. Articles providing methods information only (i.e., not reporting data) were not considered among the formal set of included articles, but were used to supplement the abstractions of the studies they referenced.

[b]Given the high volume of literature available in English-language publications and concerns about the applicability of non-English publication studies to settings in the United States, non-English articles were excluded.

Abbreviations: BDI-IA=Beck Depression Inventory-IA; BDI-II=Beck Depression Inventory-II; BPDS=Bromley Postnatal Depression Scale; CES-D=Center for Epidemiologic Studies Depression Scale; DSM-IV-TR=Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision; EPDS=Edinburgh Postnatal Depression Scale; GHQ-D=General Health Questionnaire; HADS=Hospital Anxiety and Depression Scale; HRSD=Hamilton Rating Scale for Depression; KQ=Key Question; LQ=Leverton Questionnaire; MADRS=Montgomery Asberg Depression Rating Scale; PDPI-R=Postpartum Depression Predictors Inventory-Revised; PDSS=Postpartum Depression Screening Scale; PHQ-2=Patient Health Questionnaire-2; PHQ-9=Patient Health Questionnaire-9; PRIME-MD PHQ=Primary Care Evaluation of Mental Disorders Patient Health Questionnaire; RCT=randomized controlled trial; Zung SDS=Zung Self-Rating Depression Scale

Study Selection

Using the prespecified inclusion and exclusion criteria described in Table 3, two investigators independently reviewed titles and abstracts for potential relevance to the KQs. Articles included by either reviewer underwent full-text screening. At the full-text review stage, paired researchers independently reviewed the articles and indicated a decision to “include” or “exclude” the article for data abstraction. When the two reviewers arrived at different decisions about whether to include or exclude an article, they reconciled the difference through review and discussion, or through a third-party arbitrator if needed. Full-text articles meeting our eligibility criteria were included for data abstraction. Relevant systematic review articles, meta-analyses, and methods articles were flagged for manual searching of references and cross-referencing as appropriate against the library of citations identified through electronic database searching.

For citations retrieved by searching the grey literature, the above-described procedures were modified such that a single screener initially reviewed all search results; final eligibility of citations for data abstraction was determined by duplicate screening review. All screening decisions were made and tracked in a Distiller SR database (Evidence Partners Inc., Manotick, ON, Canada).

Data Extraction

The research team created data abstraction forms and evidence table templates for abstracting data for each KQ. Based on clinical and methodological expertise, a pair of investigators was assigned to abstract data from each eligible article. One investigator abstracted the data, and the second reviewed the completed abstraction form alongside the original article to check for accuracy and completeness. Disagreements were resolved by consensus, or by obtaining a third reviewer’s opinion if consensus could not be reached. To aid in both reproducibility and standardization of data collection, researchers received data abstraction instructions directly on each form created specifically for this project within the DistillerSR database.

We designed the data abstraction forms to collect the data required to evaluate the specified eligibility criteria for inclusion in this review, as well as demographic and other data needed for determining outcomes (screening test performance characteristics, as well as intermediate, final, and adverse events outcomes). We paid particular attention to describing the details of the screening intervention that may be related to outcomes, including setting, provider, timing, and frequency of screening; patient characteristics (e.g., age, parity); and study design (e.g., randomized controlled trial [RCT] versus observational). In addition, we described comparators carefully, as treatment standards may have changed during the study period. Harms outcomes were framed to help identify adverse events (e.g., stigmatization, decreased quality of life). We also abstracted data necessary for assessing quality and applicability, as described in the *Methods Guide*[31] and *Medical Test Guide*.[32] Before the data abstraction form templates were used, they were pilot-tested with a sample of included articles to ensure that all relevant data elements were captured and that there was consistency/reproducibility between abstractors. Forms were revised as necessary before full abstraction of all included articles. Appendix B provides a detailed listing of the elements included in the data abstraction forms.

Quality Assessment of Individual Studies

We assessed the methodological quality, or risk of bias, of individual studies using the assessment instruments detailed in the *Methods Guide*[31] and *Medical Test Guide*.[32] Briefly, we assessed each study with an overall summary rating based on its adherence to well-accepted standard methodologies (e.g., the QUality Assessment for Diagnostic Accuracy Studies-2 [QUADAS-2] tool[65] for studies of diagnostic accuracy). To assess quality for studies presenting information on patient-centered intermediate, final, and adverse effect outcomes, we used a strategy to (1) classify the study design, (2) apply predefined criteria for quality and critical appraisal, and (3) arrive at a summary judgment of the study's quality. We applied criteria for each study type derived from core elements described in the *Methods Guide*. Criteria of interest for all studies included similarity of groups at baseline, extent to which outcomes were described, blinding of subjects and providers, blinded assessment of the outcome(s), intention-to-treat analysis, differential loss to followup between the compared groups or overall high loss to followup, and conflicts of interest. Criteria specific to RCTs included methods of randomization and allocation concealment. For observational studies, additional elements such as methods for selection of participants, measurement of interventions/exposures, addressing any design-specific issues, and controlling confounding were considered. To indicate the summary judgment of the quality of individual studies, we used the overall ratings of good, fair, or poor based on the study's adherence to well-accepted standard methodologies (Table 4). Studies of different designs were graded within the context of their respective designs. Thus, RCTs were graded as good, fair, or poor, and observational studies were separately graded as good, fair, or poor.

Table 4. Definitions of overall quality ratings

Quality Rating	Description
Good	A study with the least bias; results are considered valid. A good study has a clear description of the population, setting, interventions, and comparison groups; uses a valid approach to allocate patients to alternative treatments; has a low dropout rate; and uses appropriate means to prevent bias, measure outcomes, and analyze and report results.
Fair	A study that is susceptible to some bias but probably not enough to invalidate the results. The study may be missing information, making it difficult to assess limitations and potential problems. As the fair-quality category is broad, studies with this rating vary in their strengths and weaknesses. The results of some fair-quality studies are possibly valid, while others are probably valid.
Poor	A study with significant bias that may invalidate the results. These studies have serious errors in design, analysis, or reporting; have large amounts of missing information; or have discrepancies in reporting. The results of a poor-quality study are at least as likely to reflect flaws in the study design as to indicate true differences between the compared interventions.

For studies assessing screening test performance (KQs 1, 2, and 3), we used QUADAS-2[65] to assess quality. QUADAS-2 describes risk of bias in four key domains: patient selection, index test(s), reference standard, and flow and timing. The questions in each domain are rated in terms of risk of bias and concerns regarding applicability, with associated signaling questions to help with these bias and applicability judgments. Summary judgments for these studies were assigned as high risk of bias, low risk of bias, or unclear.

Data Synthesis

We began our data synthesis by summarizing key features of the included studies for each KQ. To the degree that data were available, we abstracted information on study design; patient characteristics; clinical settings; interventions; screening test performance; and intermediate, final, and adverse event outcomes.

We determined the feasibility of completing a quantitative synthesis (i.e., meta-analysis). Feasibility depended on the volume of relevant literature, conceptual homogeneity of the studies (both in terms of study population and outcomes), and completeness of the reporting of results.

We considered meta-analysis for comparisons where at least three conceptually homogenous studies reported the same patient-centered intermediate, final, or adverse effect outcome. When a meta-analysis was appropriate, we used random-effects models to synthesize the available evidence quantitatively using Comprehensive Meta-Analysis software (Version 2; Biostat, Englewood, NJ). We tested for heterogeneity using graphical displays and test statistics (Q and I^2 statistics), while recognizing that the ability of statistical methods to detect heterogeneity may be limited. For comparison purposes, we also performed fixed-effect meta-analyses. We present summary estimates, standard errors, and exact confidence intervals (CIs) in our data synthesis.

Test performance was summarized using sensitivity, specificity, and likelihood ratios. Test sensitivity describes the proportion of subjects with the disorder who have an abnormal test. Test specificity describes the proportion of subjects without the disorder who have a normal test. A likelihood ratio is a measure that may be more useful to clinicians since a simple nomogram allows posttest disorder probabilities to be readily calculated. The positive likelihood ratio (LR+) describes how many times more likely it is that an abnormal test comes from a patient with the disorder versus a patient without the disorder. The negative likelihood ratio (LR-) describes how

many times more likely it is that a normal test comes from a patient with the disorder versus a normal patient without the disorder.

When test performance studies were conceptually homogeneous, we used random-effects bivariate meta-analysis using SAS version 9.2 (SAS Institute Inc., Cary, NC) to compute summary estimates of performance.[66] A random-effects model assumes that variability is a result of sampling errors as well as the true differences between studies and provides a meta-analytic modeling approach for pooling sensitivity and specificity, while accounting for possible correlation between sensitivities and specificities of the studies included.[67] We evaluated statistical heterogeneity by inspecting forest plots and computing Q and I^2 statistics. Since the Q test is underpowered, we set the threshold for significant heterogeneity at $p<0.10$. For the I^2 test, a suggested interpretation is to assign the terms low, moderate, and high to I^2 values of 25 percent, 50 percent, and 75 percent, respectively.[68]

We anticipated that intervention effects might be heterogeneous. We hypothesized that the methodological quality of individual studies, study type, characteristics of the screening population (e.g., age, parity), and characteristics of the screening intervention (e.g., setting, provider) would be associated with the intervention effects. Where there were sufficient studies, we performed subgroup analyses and/or meta-regression analyses to examine these hypotheses. Examples of such a subgroup analysis are a comparison of effectiveness estimates for RCTs vs. observational studies, or a comparison of estimates of the association between a history of intimate partner violence and postpartum depression for cohort vs. case-control studies.

We also adapted an existing simulation model of pregnancy and neonatal outcomes[69] to estimate the balance of benefits and harms of different screening strategies based on the literature review, using the benefits and harms listed above. Because there are numerous unresolved issues about the use of quality-adjusted life years (QALYs) in the setting of maternal-child health,[70] we used the estimated likelihood of specific outcomes as the model output. Specific benefits include estimates of treated depression, false negatives, and false positives. Based on our preliminary review of the literature and discussions with the Key Informants and TEP, we expected data on harms, in particular, to be sparse. We can, however, readily estimate the number of false-positive screening test results, or total referrals for further evaluation, under different scenarios. This allows an approach which compares total tests or false-positive results as a measure of "cost" or "harm" with a measure of benefit, such as "cases of depression detected." Such an approach has been used by modelers supporting the U.S. Preventive Services Task Force (USPSTF) in making recommendations—for example, in colorectal cancer screening, where the metric was colonoscopies per cancer death prevented, or in cervical cancer screening, where the metric was colposcopies per cancer death prevented.

The model simulates pregnancy from conception through delivery and can subsequently simulate both maternal and child outcomes. Child outcomes are conditioned on gestational age at delivery and maternal race/ethnicity; both maternal and child outcomes can also easily be conditioned on maternal exposures at any point in gestation. In this context, using this model, estimates of benefits and harms can be generated for specific screening tests, at different times during and after pregnancy, for mothers and infants (and for fathers, if data are available). For example, the model could compare estimated maternal and infant outcomes from screening with a test of sensitivity X percent and specificity Y percent at 36 weeks gestation and 6 weeks postpartum, versus screening with a test of sensitivity A percent and specificity B percent at each well-child visit. The values for sensitivity and specificity (along with CIs) were derived from the literature review. The model also incorporates variability in followup and appropriate treatment

after a positive screening test result. We used probabilistic sensitivity analysis to assess overall uncertainty based on the available literature, and used a modified value-of-information approach to help prioritize future research needs.[71] Because the report found almost no evidence from which to derive estimates for longer term outcomes, we focused the analysis on estimating the number of detected cases of depression, false negative and false positive results under different scenarios of test performance, and prevalence of depression.

Strength of the Body of Evidence

We rated the strength of evidence for each KQ and outcome using the approach described in the *Methods Guide*[31,72] and *Medical Test Guide*.[32] In brief, the approach requires assessment of four domains: risk of bias, consistency, directness, and precision (Table 5).

Table 5. Strength of evidence—required domains

Domain	Rating	How Assessed
Risk of bias	Low Medium High	Assessed primarily through study design (RCT versus observational study) and aggregate study quality
Consistency	Consistent Inconsistent Unknown/not applicable	Assessed primarily through whether effect sizes are generally on the same side of "no effect" and the overall range of effect sizes
Directness	Direct Indirect	Assessed by whether the evidence involves direct comparisons or indirect comparisons through use of surrogate outcomes or use of separate bodies of evidence
Precision	Precise Imprecise	Based primarily on the size of the CIs of effect estimates

Abbreviation: CIs=confidence intervals; RCT=randomized controlled trial

Additional domains were used when appropriate, namely, strength of association (magnitude of effect) and publication bias. These domains were considered qualitatively, and a summary rating of "high," "moderate," or "low" strength of evidence was assigned after discussion by two reviewers. In some cases, high, moderate, or low ratings were impossible or imprudent to make; for example, when no evidence was available or when evidence on the outcome was too weak, sparse, or inconsistent to permit any conclusion to be drawn. In these situations, a grade of "insufficient" was assigned. This four-level rating scale consists of the following definitions:

- **High—**High confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect.
- **Moderate—**Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate.
- **Low—**Low confidence that the evidence reflects the true effect. Further research is likely to change the confidence in the estimate of effect and is likely to change the estimate.
- **Insufficient—**Evidence either is unavailable or does not permit estimation of an effect.

Applicability

We assessed applicability across our KQs using the method described in the *Methods Guide*[31,72] and *Medical Test Guide;*[32] In brief, this method uses the PICOTS format as a way to organize information relevant to applicability. Items of particular interest that may contribute to heterogeneity and impact applicability include setting (e.g. country, provider), comparator, spectrum of disease (e.g. screening population or pre-selected), patient income, race, ethnicity, parity, and partner support. We used checklists to guide the assessment of applicability (see the relevant sections of Appendix B). We used these data to evaluate the applicability to clinical practice, paying special attention to study eligibility criteria, demographic features of the enrolled population in comparison to the target population, characteristics of the intervention used in comparison with care models currently in use, and clinical relevance and timing of the outcome measures. We summarized issues of applicability qualitatively.

Peer Review and Public Commentary

The peer review process is our principal external quality-monitoring device. Nominations for peer reviewers were solicited from several sources, including the TEP and interested Federal agencies. Experts in a range of pertinent fields (obstetrics and gynecology, adult and child psychiatry, psychology, postpartum depression screening and treatment, maternal/fetal medicine, women's health, epidemiology, health services research, informed decision making, and family medicine) along with individuals representing stakeholder and user communities have been invited to provide external peer review of this draft report; AHRQ and an associate editor will also provide comments. The draft report will be posted on the AHRQ Web site for 4 weeks to elicit public comment. We will address all reviewer comments, revising the text as appropriate, and will document everything in a disposition of comments report that will be made available 3 months after the Agency posts the final report on the AHRQ Web site. We will include a list of peer reviewers submitting comments on this draft in the final report.

Results

Introduction

We begin by describing the results of our literature searches. We then provide a brief description of the included studies. The remainder of the chapter is organized by Key Question (KQ). Under each of the six KQs, we begin by listing the key points of the findings, followed by a brief description of included studies and a detailed synthesis of the evidence. We conducted quantitative syntheses where possible, as described in the Methods chapter.

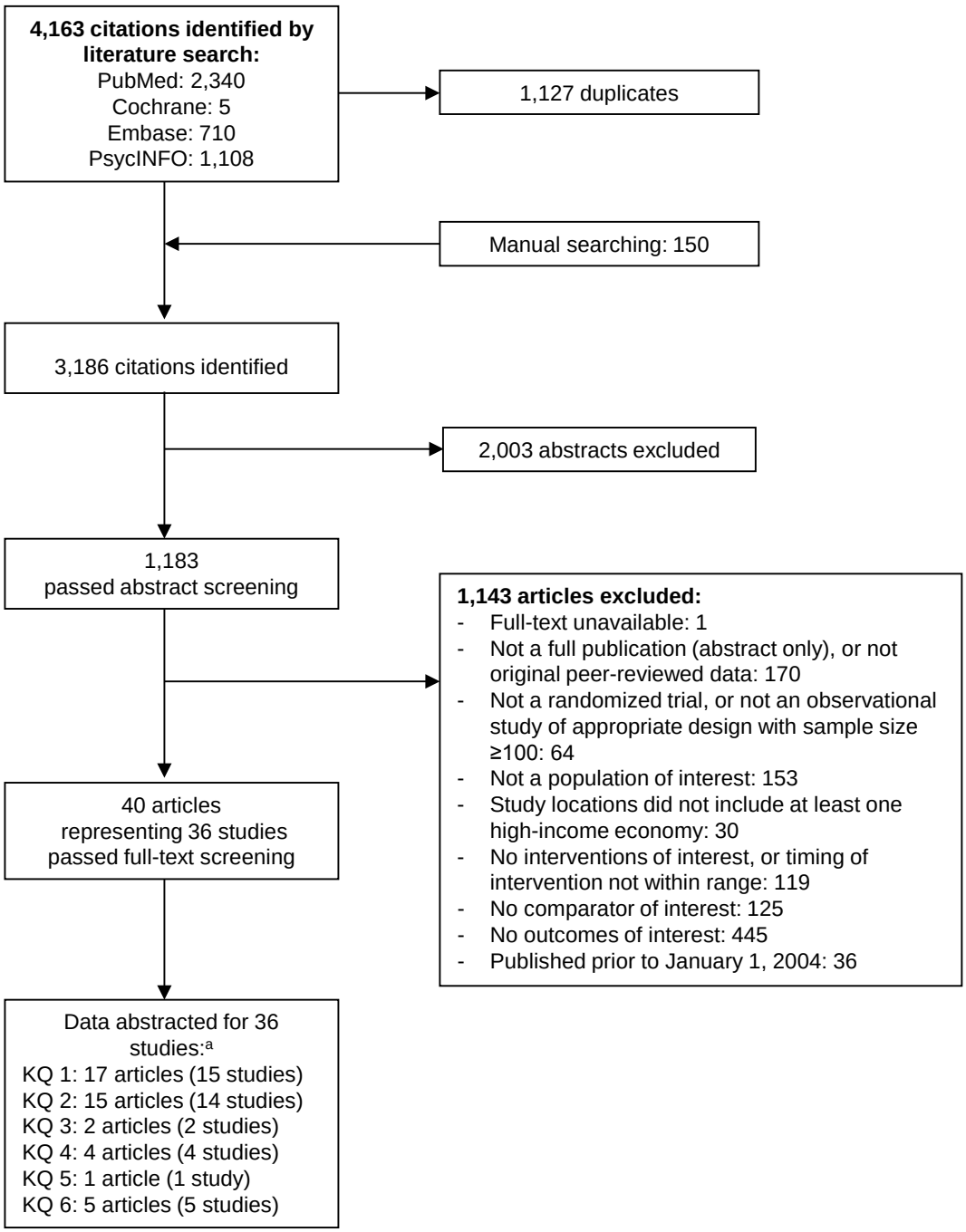
A list of abbreviations and acronyms used in this chapter is provided at the end of the report.

Results of Literature Searches

Figure 2 depicts the flow of articles through the literature search and screening process. Searches of PubMed®, Embase®, PsycINFO®, and CDSR yielded 4,163 citations, 1,127 of which were duplicate citations. Manual searching identified 150 additional citations, for a total of 3,186 citations to be screened. After applying inclusion/exclusion criteria at the title-and-abstract level, 1,183 full-text articles were retrieved and screened. Of these, 1,143 were excluded at the full-text screening stage, leaving 40 articles for data abstraction. These 40 articles described 36 unique studies. The relationship of studies to the review questions is as follows: 15 studies relevant to KQ 1, 14 studies relevant to KQ 2, 2 studies relevant to KQ 3, 4 studies relevant to KQ 4, 1 study relevant to KQ 5, and 5 studies relevant to KQ6 (some studies were relevant to more than one KQ).

Appendix C provides a detailed listing of included articles and associated publications that were used during abstraction to provide additional details on study methods. Appendix D provides a complete list of articles excluded at the full-text screening stage, with reasons for exclusion.

Figure 2. Literature flow diagram



[a]Some studies were relevant to more than one KQ.

Description of Included Studies

Overall, we included 36 studies represented by 40 publications: 15 studies were relevant to KQ 1, 14 studies to KQ 2, 2 studies to KQ 3, 4 studies to KQ 4, 1 study to KQ 5, and 5 studies to KQ 6. Studies were conducted in Europe (39%), the United States or Canada (31%), the UK (14%), Asia (8%), Australia or New Zealand (6%), and other locations (3%). Further details on

the studies included for each KQ are provided in the relevant results sections, below, and in Appendix E.

As described in the Methods chapter, we searched ClinicalTrials.gov to identify completed but unpublished studies as a mechanism for ascertaining publication bias. Our search yielded 111 trial records. A single reviewer identified six of these records as potentially relevant; three of the six had been completed at least 1 year prior to our search of the database and review of the published literature. We identified and screened publications for all three of these studies.

Two of the six identified studies are currently recruiting participants. Of these, 1 study with a target enrollment of 650 women was considered potentially applicable to KQs 4, 5, and 6. The other, with a targeted enrollment of 30 women, was considered potentially relevant to KQ 6. Upon completion, these 2 studies may provide additional evidence on the comparative benefits and/or harms of screening, or on how various factors affect the likelihood of an appropriate action after a positive screening result. One additional study with a targeted enrollment of 170 women was identified that may be relevant to KQ 6; however, that study record has not been updated since its initial entry into ClinicalTrials.gov in 2008. At that time, the study was not yet open for participant recruitment. We did not find any further information suggesting that this study has since progressed to begin enrolling. In summary, our search of ClinicalTrials.gov did not find evidence for completed but unpublished studies relevant to our KQs.

Key Question 1. Performance Characteristics of Screening Instruments

KQ 1: This question has two parts:

a. What are the sensitivity and specificity of currently available screening instruments for detecting postpartum depression, and how do these translate into the likelihood of false-negative and false-positive results in different populations and settings?

b. Are there clinically relevant differences in the ability of currently available screening instruments to correctly identify specific signs or symptoms of depression (e.g., suicidal ideation)?

Key Points

- Studies of individual screening tests rarely used the same threshold, preventing meaningful quantitative comparison or synthesis.
- For any given screening test, sensitivity was generally higher and specificity lower as the threshold for postpartum depression on the screening test was lowered.
- Precision was better for specificity estimates than for sensitivity estimates.
- Both sensitivity and specificity generally were in the 80-90% range for most screening tests across studies.
- In one study at low risk of bias, a two-question screen had a sensitivity of 100 percent (95% confidence interval [CI], 93.3 to 100%), with a specificity of 44.0 percent (95% CI, 39.5 to 48.8%), suggesting that it is possible to use an initial simple step for selecting patients for more specific screening instruments.

- In one study at low risk of bias, the 24-item Leverton Questionnaire had higher sensitivity (95.2%; 95% CI, 90.4 to 98.1%) and specificity (91.3%; 95% CI, 88.4 to 93.7%) than was generally observed for other screening instruments, but we did not identify any confirmatory studies in a U.S. population.
- We did not identify any studies comparing the ability of screening instruments to correctly identify specific signs or symptoms of depression. One study found moderate agreement between suicidal ideation on the Edinburgh Postnatal Depression Scale (EPDS) and suicidal ideation on the diagnostic instrument, but suicidal ideation was not predictive of response to treatment.
- For the EPDS and Postpartum Depression Screening Scale (PDSS), the two most commonly studied screening tools, the overall strength of evidence was moderate that both sensitivity and specificity are generally in the 80 to 90 percent range; the evidence was insufficient for other screening tests. The evidence was also insufficient to determine if there are any clinically meaningful differences in test characteristics between individual screening tests.

Description of Included Studies

We identified 15 studies that met the inclusion criteria for KQ 1.[73-87] All confirmed the diagnosis of depression using a validated clinical interview or diagnostic instrument in screen positives and all or a sample of screen negatives. Four studies[75,76,81,85] were performed in the United States, five[78,80,84,86,87] in Europe, three[74,79,83] in the UK, and one each in Australia,[73] Asia,[82] and Canada.[77] Six were judged to have a high risk of biased results;[73,74,77,79,84,86] the remainder were judged to be at low risk.

Because no more than two studies provided results for the same test at the same threshold, we did not perform meta-analyses. Below, we present and discuss the results of the studies for each screening test qualitatively, then present the results for the three studies where two or more screening tests were directly compared. Only one study[83] was relevant to KQ 1b.

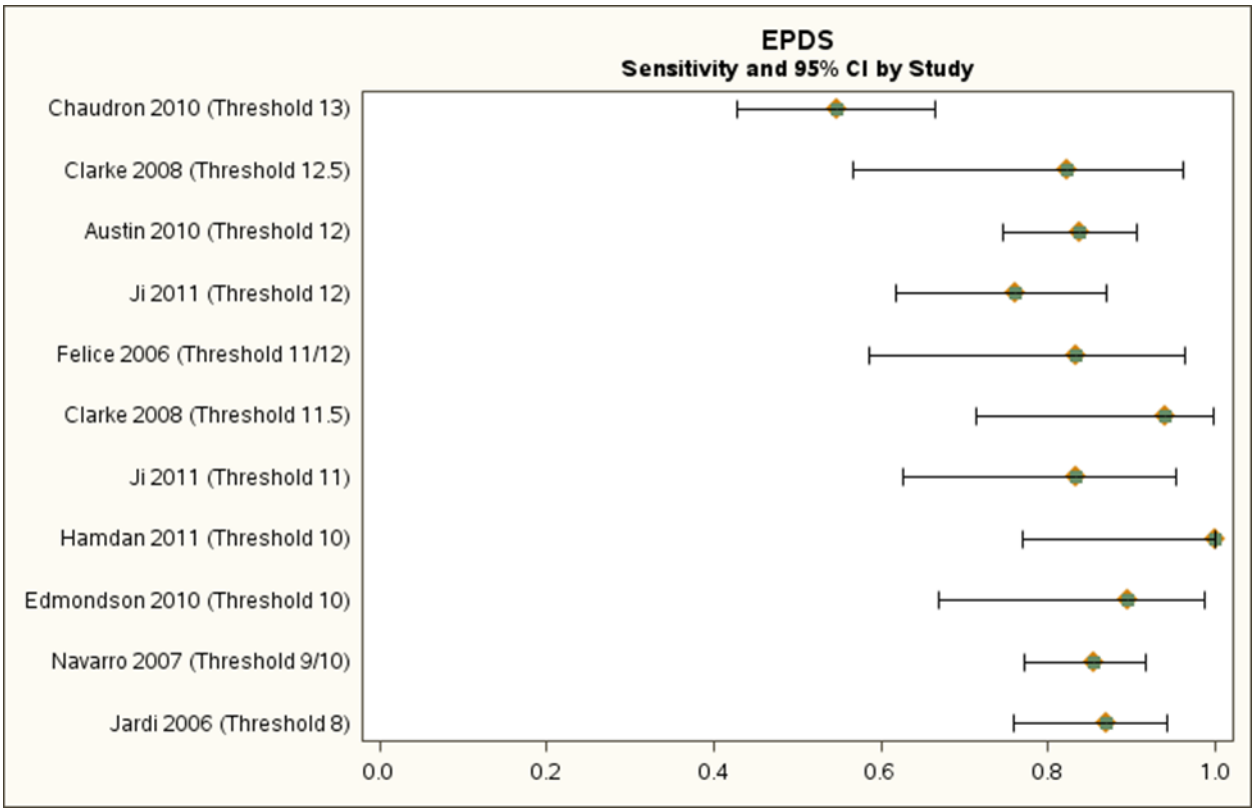
Ten studies provided data on the Edinburgh Postnatal Depression Scale (EPDS), four on the Postpartum Depression Screening Scale (PDSS), two on the Beck Depression Inventory (BDI), and one each on the Antenatal Risk Questionnaire, the 17- and 21-Item Hamilton Rating Scale for Depression (HRSD-17 and HRSD-21), the Leverton Questionnaire, and the Patient Health Questionnaire (PHQ).

Detailed Synthesis

Edinburgh Postnatal Depression Scale (EPDS)

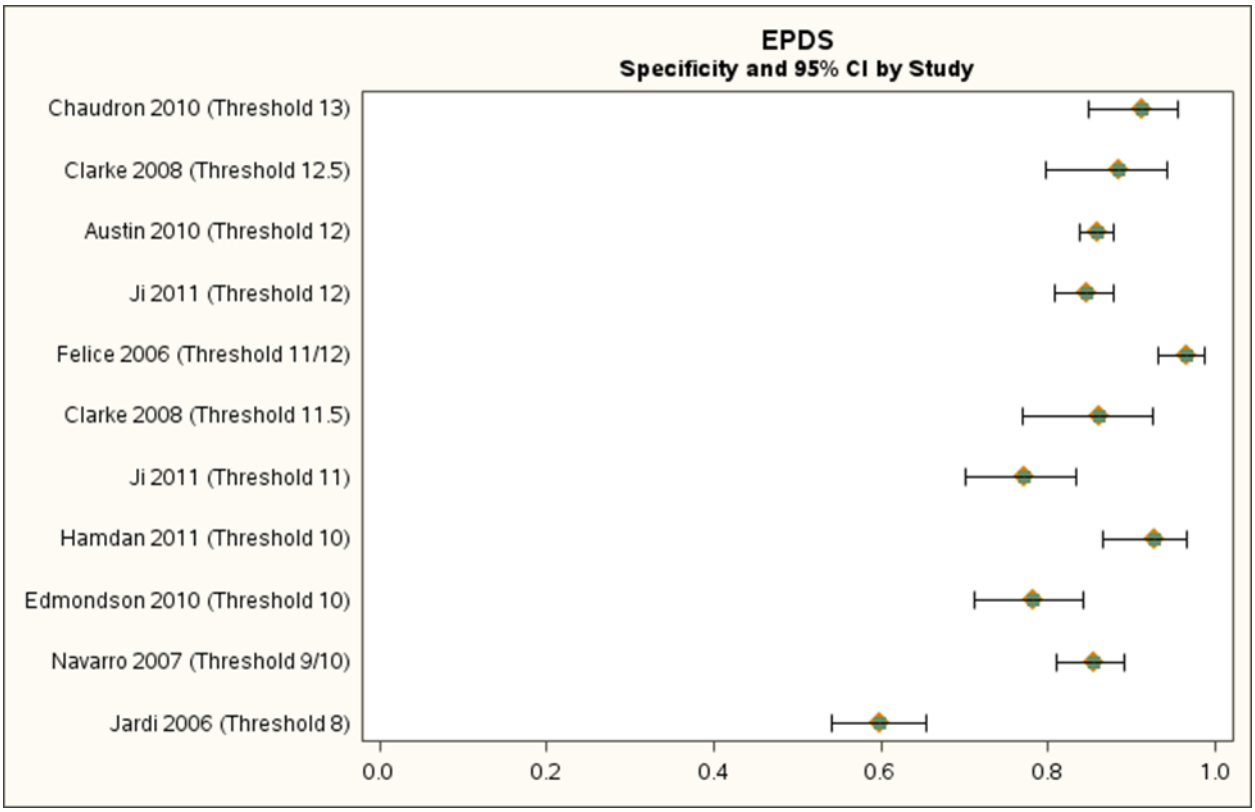
Ten studies provided sensitivity and specificity data for major depression for the EPDS.[74,76,77,79,80,82-86] Studies varied in the threshold used to define a positive screening test. There was a clear trend towards increasing sensitivity and decreasing specificity as the threshold value decreased (Figures 3 and 4). For sensitivity, confidence intervals were wide and overlapped, except for the studies which used thresholds of 8[84] and 13.[76] Even though, as expected, confidence intervals were considerably narrower for specificity, there was again considerable overlap across thresholds from 10 through 12.

Figure 3. Sensitivity of the EPDS at various thresholds



Abbreviations: CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale

Figure 4. Specificity of the EPDS at various thresholds

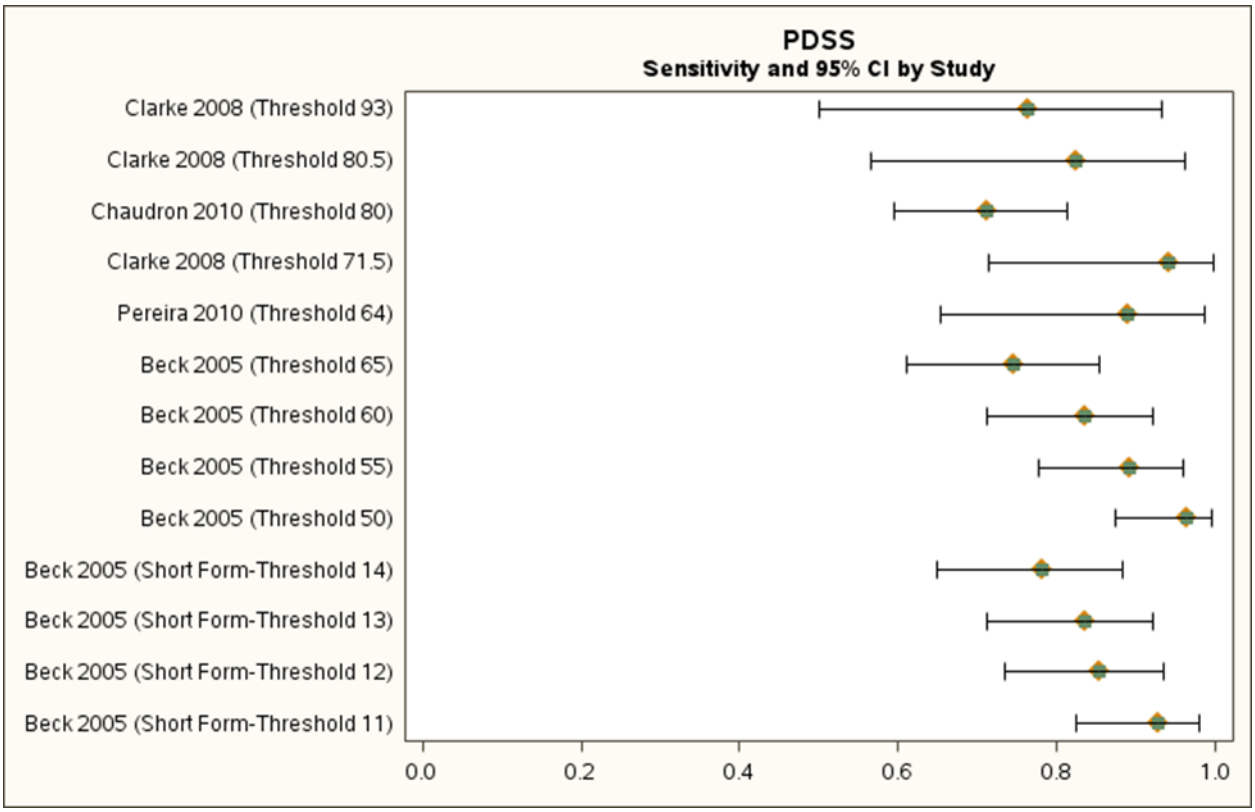


Abbreviations: CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale

Postpartum Depression Screening Scale (PDSS)

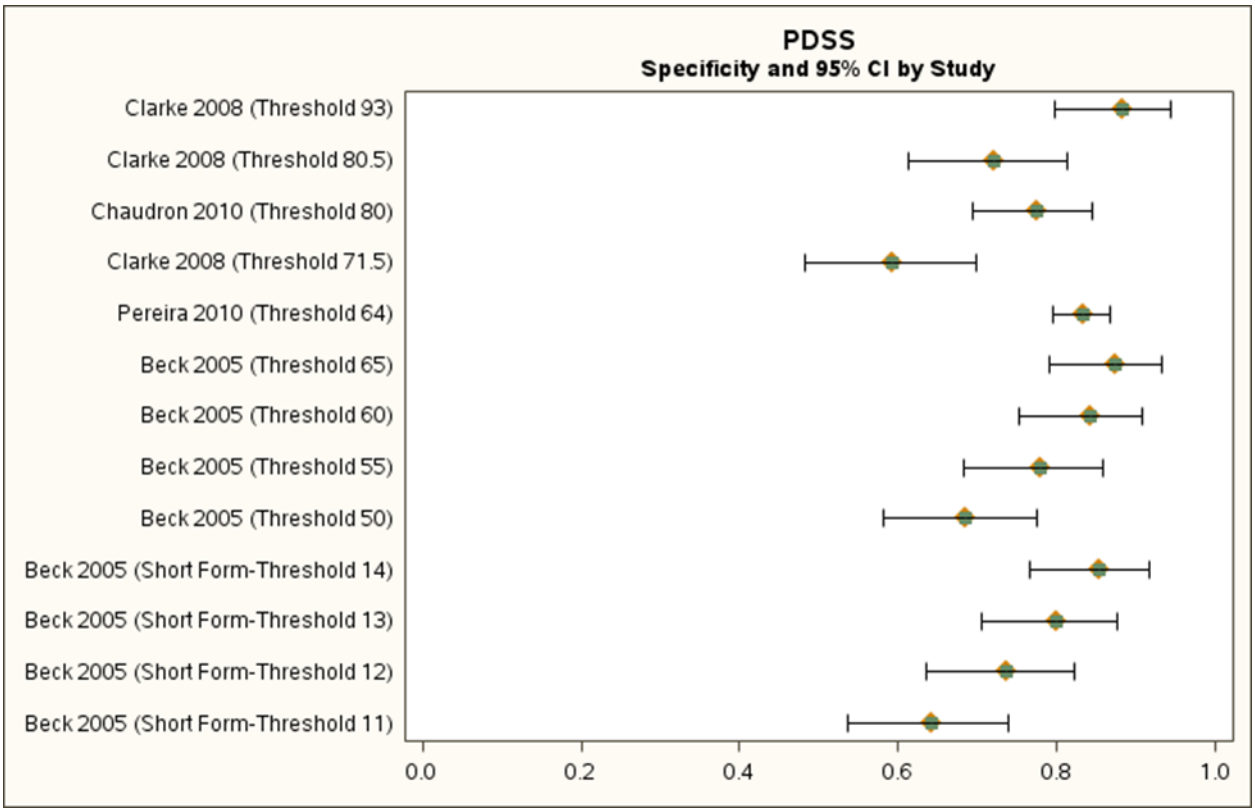
Four studies provided sensitivity and specificity data for the PDSS across a range of thresholds.[75-77,87] Of note, Beck et al.[75] was a validation study of a long- and short-form Spanish version of the PDSS in a U.S. Latina population and presented results primarily for combined major and minor depression. Figures 5 and 6 depict results for the Beck study[75] (major and minor depression combined), and for five studies where the outcome was major depression alone; they also indicate whether the long- or short-form PDSS was used. As with EPDS, confidence limits are wider for sensitivity than for specificity, and there is a clear trend towards increasing sensitivity and decreasing specificity as thresholds decreased. Qualitatively, the values for sensitivity and specificity at a given threshold were similar, with sensitivities between 80 and 90 percent associated with specificities in the same range.

Figure 5. Sensitivity of the PDSS at various thresholds



Abbreviations: CI=confidence interval; PDSS=Postpartum Depression Screening Scale

Figure 6. Specificity of the PDSS at various thresholds



Abbreviations: CI=confidence interval; PDSS=Postpartum Depression Screening Scale

Other Tests

There were only one or two studies for the other screening tests identified in the review. Table 6 summarizes sensitivity/specificity results for these studies.

Table 6. Test characteristics for postpartum depression screening instruments other than EPDS or PDSS

Screening Test	Study	Sensitivity (95% CI)	Specificity (95% CI)
Antenatal Risk Questionnaire	Austin, 2011[73]	78.2% (65.0 to 88.2%)	47.1% (40.3 to 53.9%)
BDI	Ji, 2011[85]	82.8% (73.9 to 90.0%	82.1% (79.2 to 85.0%)
BDI-1A	Csatordai, 2009 [78]	91.8% (86.1% to 95.7%)	91.5% (88.7 to 93.9%)
HRSD-17	Ji, 2011[85]	81.0% (65.9 to 91.4%)	80.9% (76.4 to 85.0%)
HRSD-21	Ji, 2011[85]	81.0% (65.9 to 91.4%	75.7% (70.8 to 80.1%)
Leverton Questionnaire	Csatordai, 2009[78]	95.2% (90.4 to 98.1%)	91.3% (88.4 to 93.7%)
PHQ-2 yes/no	Gjerdingen, 2009[81]	100% (93.3% to 100%)	44.0% (39.5 to 48.8%)
PHQ-2 Likert	Gjerdingen, 2009[81]	84.4% (70.5 to 93.5%)	78.7% (74.7 to 82.4%)
PHQ-9-simple	Gjerdingen, 2009[81]	82.2% (68.0 to 92.0%)	83.8% (80.1 to 87.0%)

Screening Test	Study	Sensitivity (95% CI)	Specificity (95% CI)
PHQ-9 complex	Gjerdingen, 2009[81]	66.7% (52.9 to 80.4%)	91.5% (89.0 to 94.1%)

Abbreviations: BDI=Beck Depression Inventory; BDI-IA=Beck Depression Inventory-IA; EPDS=Edinburgh Postnatal Depression Scale HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression; PHQ-2=2-Item Patient Health Questionnaire; PHQ-9=9-Item Patient Health Questionnaire; PDSS=Postpartum Depression Screening Scale

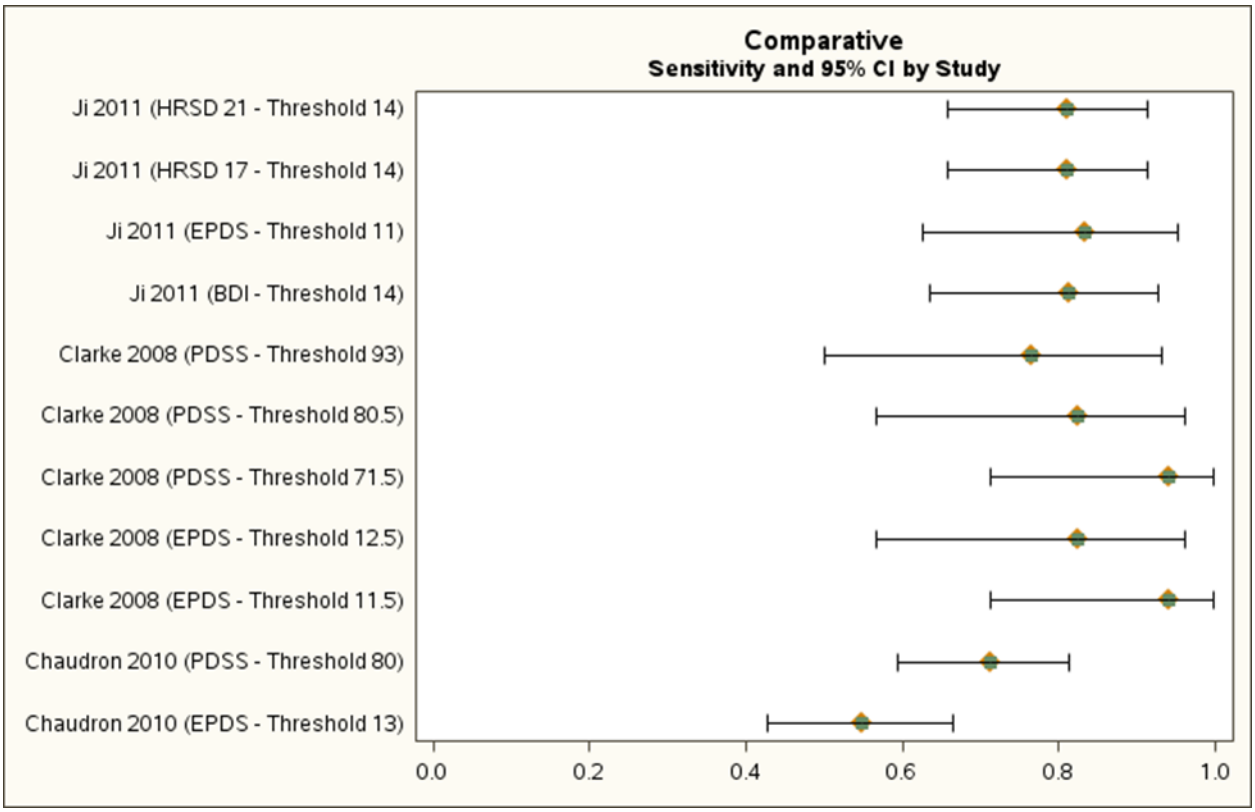
Results for these other tests were generally consistent with those for the EPDS and PDSS: precision of the estimate was greater for specificity than for sensitivity, and both sensitivity and specificity point estimates were generally in the 80–90 percent range. There were several exceptions to these general observations. The Antenatal Risk Questionnaire had a sensitivity at the low range of those of the other tests but a lower specificity. A screen consisting of two questions ("During the past month, have you often been bothered by feeling down, depressed, or hopeless?" and "During the past month, have you often been bothered by having little interest or pleasure in doing things?") had a sensitivity of 100 percent if the response to either question was yes, but low specificity.[81] Finally, the BDI and the 25-item Leverton Questionnaire had both sensitivity and specificity above 90 percent in a Hungarian validation study.[81] This study was rated as low risk of bias, and had 1,552 subjects.

Within-Study Comparisons

Three studies compared different screening instruments in the same population (Figures 7 and 8).[76,77,85] The largest[85] compared the EPDS, BDI, HSRD-17, and HSRD-21 at multiple time points across pregnancy. Performance characteristics were similar for all the tests when performed from 8 weeks to 6 months postpartum, which are the time points depicted in Figures 7 and 8 (the similarity in characteristics across tests was also seen in the early postpartum period—differences within tests by timing are discussed under KQ 6).

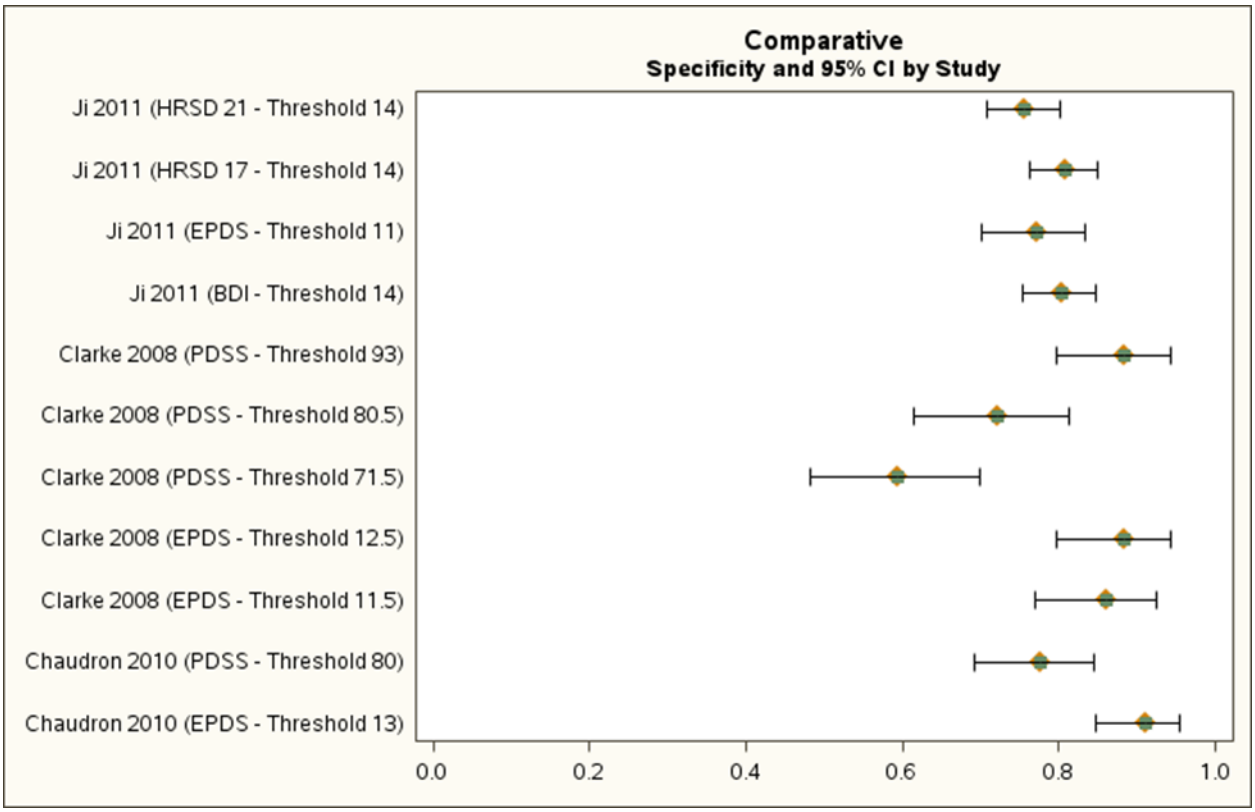
The two studies comparing the EPDS and PDSS[76,77] found slightly lower sensitivity but higher specificity for the EPDS, depending on the threshold, but there was considerable overlap in confidence limits, especially for sensitivity.

Figure 7. Comparative sensitivity of various screening instruments



Abbreviations: BDI=Beck Depression Inventory; EPDS=Edinburgh Postnatal Depression Scale HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression; PDSS=Postpartum Depression Screening Scale

Figure 8. Comparative specificity of various screening instruments



Abbreviations: BDI=Beck Depression Inventory; EPDS=Edinburgh Postnatal Depression Scale HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression; PDSS=Postpartum Depression Screening Scale

Specific Signs and Symptoms

We identified one study performed in Scotland which evaluated the suicidal ideation component of the EPDS.[83] Howard et al performed a nested cohort study within an RCT for postpartum depression. The prevalence of suicidal ideation on the EPDS at 6–8 weeks was 9.0 percent (95% CI, 8.3 to 10.1%), with 4 percent (95% CI, 3.2 to 4.4%) reporting having thoughts of harming themselves sometimes or quite often. Agreement between suicidality in the EPDS versus the diagnostic instrument used in this study (CIS-R) was only moderate, with a kappa of 0.66; 68 percent of those with some suicidal ideation on the EPDS would have been defined as suicidal using CIS-R criteria. Suicidal ideation at baseline did not correlate with any outcome in the trial at 18 weeks postpartum, including SF-12 physical and mental scores and repeat EPDS score, even after adjustment for differences between those with and without suicidal ideation (younger, unmarried, unemployed or with an unemployed partner, and worse quality of relationship).

Key Question 2. Effect of Individual Subject Factors on Screening Performance

KQ 2: This question has two parts:

a. Are there individual factors (age, race, parity [number of live births], history of mood disorders, history of intimate partner violence, perinatal outcomes, cultural factors) that affect the baseline risk of postpartum depression and, therefore, the subsequent positive and negative predictive values of screening instruments?

b. Are there validated predictive models or algorithms based on such factors that would improve the performance of screening instruments?

Key Points

- The positive and negative predictive values of screening for postpartum depression are affected by the prevalence of depression; screening women who are at higher risk would improve the positive predictive value.
- Maternal age and socioeconomic status were generally not associated with risk of postpartum depression; maternal unemployment increased the risk of postpartum depression in one study. The overall strength of evidence was low.
- Complications of pregnancy, including preterm birth, low birthweight, and fetal abnormalities are associated with an increased risk of postpartum depression. Parity was not consistently associated with postpartum depression. One study found no significant effect of parity on test characteristics for screening tests. The overall strength of evidence was moderate.
- Chronic medical conditions predating pregnancy may increase the risk of postpartum depression. The strength of evidence was low.
- Past history of depression or anxiety, whether or not associated with a previous pregnancy, consistently increases the risk of postpartum depression. The strength of evidence was moderate.
- Poor relationship quality and poor social support consistently increase the risk of postpartum depression. The strength of evidence was moderate.
- We did not identify any studies of clinical predictive models or algorithms (comparable to the Gail model for breast cancer risk) for improving the performance of screening instruments.

Description of Included Studies

We identified 15 articles describing 14 unique studies that met the inclusion criteria for KQ 2.[84,85,88-100] (The 2005[95] and 2008[94] publications by Chee et al. described results for the same study population.) Three were from the United States,[85,88,91] seven were from Europe,[84,89,96-100] two were Asian,[93,95] and there was one study each from the UK,[90] Australia,[93] and Israel.[92] Five were judged to be of good quality,[89,90,93,96,99] with the rest being fair. We did not identify any studies relevant to KQ 2b. One study judged to be at high risk of bias[85] did not provide an

estimate of the association between parity and postpartum depression, but did provide separate estimates of screening test sensitivity and specificity stratified by parity.

Because of the inconsistency in how specific risk factors were described in the studies, we were unable to perform quantitative synthesis of the results.

Detailed Synthesis

Because we were unable to perform meta-analyses for any of the risk factors, we summarize the results in a series of tables below. Unless otherwise noted, results in the tables are presented for the final multivariate analysis (usually logistic regression) presented in each paper, and represent the results for an outcome of major depression. Some reported associations were not amenable to display in the table and are discussed separately in the text.

Among potential maternal demographic risk factors (Table 7), no statistically significant association was found between maternal age, education, income, or type of employment. One study[84] did, however, find a significant association between maternal unemployment and postpartum depression (OR 2.8; 95% CI, 1.1 to 4.9).

Table 7. Maternal demographic risk factors for postpartum major depression

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
Maternal age				
Akincigil, 2010[88]	Maternal age ≤21 years	Referent	-	**Maternal age**, marital status, history of domestic violence, disagreement about pregnancy, race, education, poverty, birthweight, number of children, social support, self-rated health status, prenatal tobacco/alcohol use
	Maternal age 22–24	1.08	NS	
	Maternal age ≥ 25	0.81	NS	
Boyce, 2005[93]	Maternal age (linear variable)	0.96	0.88 to 1.05	**Maternal age**, education, baseline EPDS score, family history, past history of depression, vulnerable personality, low organized/responsive, dissatisfaction with social support, dissatisfaction with partner, worsening relationship, one or more other life events
Education/income/employment				
Akincigil, 2010[88]	Education: Less than HS	Referent	-	Maternal age, marital status, history of domestic violence, disagreement about pregnancy, race, **education**, poverty, birthweight, number of children, social support, self-rated health status, prenatal tobacco/alcohol use
	Education: HS or equivalent	1.02	NS	
	Education: Some college	1.23	NS	
	Education: College or more	1.15	NS	
Boyce, 2005[93]	Education (not specified)	1.04	0.84 to 1.27	Age, **education,** baseline EPDS score, family history, past history of depression, vulnerable personality, low organized/responsive, dissatisfaction with social support, dissatisfaction with partner, worsening relationship, one or more other life events

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
Akincigil, 2010[88]	Income/poverty ratio: <100%	Referent	-	Maternal age, marital status, history of domestic violence, disagreement about pregnancy, race, education, **poverty,** birthweight, number of children, social support, self-rated health status, prenatal tobacco/alcohol use
	Income/poverty ratio: 100–300%	1.17	NS	
	Income/poverty ratio: >300%	0.82	NS	
Andersson, 2006[89]	Type of employment: Laborer	Referent	-	Age, **socioeconomic status,** smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, 1st trimester BMI
	Type of employment: Professional	1.09	0.54 to 2.23	
Jardri, 2006[84]	Unemployed	2.8	1.1 to 4.9	History of postpartum depression, history of depression, preterm birth, stopping breast feeding, multiple gestation, postpartum complications, **employment status**

Abbreviations: BMI=body mass index; CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale; HS=high school; NR=not reported; NS=not statistically significant; OR=odds ratio; RR=relative risk

Table 8 summarizes the available data on potential risk factors relating to obstetric history. As shown there, having a preterm[84] or very low birthweight[88] baby were both significantly associated with postpartum depression. In another study not shown in Table 8,[100] having a termination of pregnancy for a severe fetal malformation or chromosomal abnormality in the 2nd or 3rd trimester was associated with a significant risk of depression 14 months after the event compared with women with healthy children, with better social support reducing the risk of depression; there was no control group of women with similar fetal abnormalities who did not undergo termination. As shown in Table 8, higher parity was significantly associated with increased risk of depression in one study,[88] with a positive but nonsignificant association in another.[89] One study[88] reported a significant increase in risk of depression with prenatal smoking, while another[89] reported a nonsignificant decrease in risk among smokers.

Table 8. Obstetric history risk factors for postpartum depression

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
Akincigil, 2010[88]	Very low birthweight infant	1.63	P<0.1	Maternal age, marital status, history of domestic violence, disagreement about pregnancy, race, education, poverty, **birthweight,** number of children, social support, self-rated health status, prenatal tobacco/alcohol use
Jardri, 2006[84]	Delivery prior to 37 weeks	4.5	1.4 to 14.6	History of postpartum depression, history of depression, **preterm birth**, stopping breast feeding, multiple gestation, postpartum complications, employment status
Akincigil, 2010[88]	Parity: ≥3 children	Referent	-	Maternal age, marital status, history of domestic violence, disagreement about pregnancy, race, education, poverty, birthweight**, number of children, s**ocial support, self-rated health status, prenatal tobacco/alcohol use
	Parity: 1–2 children	0.79	P<0.05	
Andersson, 2006[89]	Parity: Nulliparous	Referent	-	Age, socioeconomic status, smoking status, snuff use, **parity,** alcohol use, chronic disease, history of psychiatric disorder, 1st trimester BMI
	Parity: Multiparous	1.14	0.55 to 2.55	
Akincigil, 2010[88]	Prenatal tobacco use	1.23	P<0.1	Maternal age, marital status, history of domestic violence, disagreement about pregnancy, race, education, poverty, birthweight, number of children, social support, self-rated health status, **prenatal tobacco/alcohol use**
Andersson, 2006[89]	Smoker	0.28	0.03 to 2.45	Age, socioeconomic status, **smoking status**, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, 1st trimester BMI
Akincigil, 2010[88]	Prenatal alcohol use	1.14	NS	Maternal age, marital status, history of domestic violence, disagreement about pregnancy, race, education, poverty, birthweight, number of children, social support, self-rated health status, **prenatal tobacco/alcohol use**

Abbreviations: BMI=body mass index; CI=confidence interval; NS=not statistically significant; OR=odds ratio; RR=relative risk

One study[85] estimated sensitivity and specificity for the BDI, EPDS, HSRD-17, and HSRD-21 separately based on timing of screening (discharge to 8 weeks postpartum vs. 8 weeks to 6 months postpartum) and by gravidity (primigravid vs. multigravid) (Table 9). Although both the BDI and EPDS had higher sensitivity in primigravidas during both time periods, confidence intervals were wide and overlapped.

Table 9. Sensitivity and specificity of screening instruments by timing of screening and gravidity

Instrument, Timing, and Gravidity	Sensitivity			Specificity		
	Point Estimate	Lower 95% CI	Upper 95% CI	Point Estimate	Lower 95% CI	Upper 95% CI
Discharge to 8 weeks						
BDI						
Primigravid	100.0%	54.1%	100.0%	68.0%	56.2%	78.3%
Multigravid	88.5%	69.9%	97.6%	82.2%	76.5%	87.1%
EPDS						
Primigravid	100.0%	29.2%	100.0%	73.2%	57.1%	85.8%
Multigravid	81.0%	58.1%	94.6%	78.9%	70.6%	85.7%
HRSD-17						
Primigravid	75.0%	34.9%	96.8%	95.2%	88.3%	98.7%
Multigravid	82.4%	65.5%	93.2%	78.5%	72.9%	83.4%
HRSD-21						
Primigravid	75.0%	34.9%	96.8%	91.7%	83.6%	96.6%
Multigravid	85.3%	68.9%	95.1%	71.7%	65.7%	77.2%
8 weeks to 6 months						
BDI						
Primigravid	86.7%	59.5%	98.3%	82.5%	76.5%	87.5%
Multigravid	80.7%	70.6%	88.6%	83.3%	79.9%	86.4%
EPDS						
Primigravid	87.5%	47.4%	99.7%	90.6%	84.1%	95.0%
Multigravid	76.2%	60.6%	88.0%	80.6%	75.7%	84.8%
HRSD-17						
Primigravid	90.5%	69.6%	98.8%	82.2%	76.5%	87.0%
Multigravid	77.2%	67.3%	85.3%	82.4%	79.2%	85.3%
HRSD-21						
Primigravid	85.7%	63.7%	97.0%	77.6%	71.5%	83.0%
Multigravid	80.4%	70.9%	88.0%	77.5%	74.0%	80.7%

Abbreviations: BDI=Beck Depression Inventory; CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression

Among potential general medical history risk factors (Table 10), fair/poor self-reported health status[88] and a history of chronic illness outside of pregnancy[89] both increased the risk of postpartum depression over two-fold. One small study not shown in Table 10 found a significant association between maternal epilepsy and postpartum depressive symptoms;[98] however, the study was underpowered to detect the potential impact of antiepileptic drugs on affective symptoms.

Table 10. General medical history risk factors for postpartum depression

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
Akincigil, 2010[88]	Self-rated health status: Great/very good	Referent	-	Maternal age, marital status, history of domestic violence, disagreement about pregnancy, race, education, poverty, birthweight, number of children, social support, **self-rated health status**, prenatal tobacco/alcohol use
	Self-rated health status: Good	1.14	NR	
	Self-rated health status: Fair/poor	2.15	P<0.05	
Andersson, 2006[89]	History of chronic illness	2.71	2.36 to 19.14	Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, **chronic disease**, history of psychiatric disorder, 1st trimester BMI
Andersson, 2006[89]	1st trimester BMI: 18.5–24.9	Referent	-	Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, **1st trimester BMI**
	1st trimester BMI: 25–29.9	2.17	NS	
	1st trimester BMI: ≥30	No cases	NR	

Abbreviations: BMI=body mass index; CI=confidence interval; NR=not reported; NS=not statistically significant; OR=odds ratio; RR=relative risk

Past history of depression or anxiety, including both postpartum and before pregnancy, were consistently associated with an increased risk of postpartum depression, with odds ratios well above 2.0 (Table 11). Two studies also found that certain personality traits (neuroticism, vulnerability, low organization) were risk factors for depression.[93,99]

Table 11. Psychiatric history risk factors for postpartum major depression

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
History of depression				
Bloch, 2005[92]	History of premenstrual dysphoric disorder	NR	0.048	Postpartum mood, **history of depression,** history of mood symptoms while using oral contraceptives, EPDS score, **history of premenstrual dysphoric disorder**
	History of major depressive disorder	NR	0.002	
Chee, 2005[95]	History of depression	4.91	1.08 to 22.3	Education, marital dissatisfaction, confinement, low instrumental support,[a] history of depression
Garcia-Esteve, 2008[96]	History of depression	3.67	1.63 to 8.27	Family caregiver role, poor partner relationship, low social support during pregnancy
Jardri, 2006[84]	History of postpartum depression	4.3	1.7 to 10.9	History of postpartum depression, history of depression, preterm birth, stopping breast feeding, multiple gestation, postpartum complications
	History of depression	4.4	2.2 to 9.0	
Verkerk, 2005[99]	History of depression	3.08	1.10 to 8.63	Covariates not reported; results presented here for 3 months postpartum, similar findings for 6 and 12 months
	Family history of depression	1.60	0.67 to 3.85	
	Depression during pregnancy	2.10	0.82 to 5.37	

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
Other psychiatric disorders				
Andersson, 2006[89]	History of psychiatric disorder	6.72	2.36 to 19.14	Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, 1st trimester BMI
Boyce, 2005[93]	Past psychiatric history	2.74	0.60 to 12.45	Age, education, baseline EPDS score, family history, past history of depression, vulnerable personality, low organized/responsive, dissatisfaction with social support, dissatisfaction with partner, worsening relationship, one or more other life events
	Vulnerable personality	2.82	1.06 to 7.45	
	Low organized/responsive personality	3.69	1.26 to 10.8	
Mauri, 2010[97]	Any anxiety disorder	2.4	1.1 to 5.7	Other covariates not specified
	Panic disorder	8.0	2.75 to 23.3	
	Social phobia	7.65	2.64 to 22.2	
Verkerk, 2005[99]	High neuroticism/high introversion	3.08	1.10 to 8.63	Covariates not reported; results presented here for 3 months postpartum, similar findings for 6 and 12 months
	High neuroticism/low introversion	1.58	0.51 to 4.93	

Abbreviations: BMI=body mass index; CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale; NR=not reported; OR=odds ratio; RR=relative risk

Finally, although studies used a variety of different scales to measure the effect of relationship quality and social support on risk of depression, and were conducted in a wide range of settings ranging from urban United States to Singapore, the qualitative results were consistent: postpartum depression was significantly more common among women in poorer quality relationships (or no relationship), and among women with poor social support (Table 12).

Table 12. Relationship and social support risk factors for postpartum major depression

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
Marital status				
Akincigil, 2010[88]	Married	Referent	-	Maternal age, **marital status,** history of domestic violence, disagreement about pregnancy, race, education, poverty, birthweight, number of children, social support, self-rated health status, prenatal tobacco/alcohol use
	Cohabiting	0.98	NS	
	Visiting	0.97	NS	
	No relation	0.64	P<0.1	
Andersson, 2006[89]	Single (compared with married)	26.4	4.14 to 168.3	Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, 1st trimester BMI, **marital status**

Study	Risk Factor	RR/OR	95% CI or P Value	Variables Included in Multivariate Analysis
Other factors				
Akincigil, 2010[88]	History of violence/abuse	1.36	NS	Maternal age, marital status, **history of domestic violence, disagreement about pregnancy**, race, education, poverty, birthweight, number of children, social support, self-rated health status, prenatal tobacco/alcohol use
	Better relationship quality	0.89	p<0.05	
	Disagreement about pregnancy	1.41	p<0.05	
Barnes, 2009[90]	More social support	0.89	0.80 to 0.98	Education, occupation, age, ethnicity, marital status, number of children, stress, **social support**, depression at 2 months
Boyce, 2005[93]	Dissatisfaction with partner	1.38	0.23 to 8.19	Age, education, baseline EPDS score, family history, past history of depression, vulnerable personality, low organized/responsive, dissatisfaction with social support, dissatisfaction with partner, worsening relationship, one or more other life events
	Worsening relationship	2.45	0.78 to 6.47	
Chee, 2005[95]	Marital dissatisfaction	9.42	2.19 to 40.52	Education, marital dissatisfaction, confinement, low instrumental support,[a] history of depression
	Negative "confinement" experience (restricted activities per cultural norms in different ethnic communities)	19.41	2.03 to 185.5	
	Low instrumental support[a]	23.43	3.68 to 149.16	
Chee, 2008[94]	Less emotional support	1.92	1.12 to 3.68	Education, marital dissatisfaction, confinement, low instrumental support,[a] history of depression
	3 or more nonscheduled pediatric visits	2.87	1.41 to 5.85	
Garcia-Esteve, 2008[96]	Family caregiver role	4.39	1.10 to 17.4	Family caregiver role, poor partner relationship, low social support during pregnancy
	Poor partner relationship	4.24	1.38 to 13.05	
	Low social support	4.06	1.47 to 11.21	

Abbreviations: BMI=body mass index; CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale; NR=not reported; NS=not statistically significant; OR=odds ratio; RR=relative risk

[a]Instrumental support is the provision of financial assistance, material goods, or services, also called "tangible support."

Key Question 3. Effect of Testing Variables on Screening Performance

Are the performance characteristics (sensitivity, specificity, predictive values) of screening instruments affected by:

a. Timing (prenatal, peripartum, or at various times in the first postpartum year) and frequency of screening?
b. Setting (prenatal visit, hospital/birthing center/home, postpartum maternal visit, or well-child visit)?
c. Provider (obstetrician, midwife, pediatrician, family practitioner, other health provider)?

Key Points

- Screening instrument performance characteristics vary by timing of administration, but the absolute difference in sensitivity and specificity across different time points is relatively small.
- Screening for postpartum depression in the immediate postpartum period, e.g. within the first week postpartum, likely identifies only those women at highest risk of developing depression and misses those with a slower onset of symptomatology.
- When screening for depression within the first 6 weeks postpartum, the Edinburgh Postnatal Depression Scale (EPDS), 17-Item Hamilton Rating Scale for Depression (HRSD-17), 21-Item Hamilton Rating Scale for Depression (HRSD-21), and Beck Depression Inventory (BDI) appear to have equivalent performance when using optimal instrument-dependent cutoffs.

Description of Included Studies

Two studies met the inclusion criteria for KQ 3a.[85,101] No studies were identified that met the inclusion criteria for KQ 3b or KQ 3c.

Of the two KQ 3a studies, one[85] was U.S. federally funded. Women participating in this prospective investigation of maternal mental illness at a single academic center were enrolled prior to 28 weeks gestation and followed through 6 months postpartum. Participants completed the EPDS, BDI, HRSD-17, and HRSD-21 during six perinatal windows: preconception, first trimester, second trimester, third trimester, early postpartum (0–6 weeks), and later postpartum (7–26 weeks). The diagnosis of depression was confirmed by the Mood Module of the Structured Clinical Interview for Depression (SCID).

The second study[101] was funded by a combination of industry and philanthropic monies. This single-center prospective study in Dublin, Ireland, enrolled women during the immediate postpartum period to determine if the EPDS, administered prior to hospital discharge, was predictive of depression at 6 weeks postpartum. Nine hundred fifty-one enrolled women completed the EPDS at 3–5 days postpartum with planned followup at 6 weeks postpartum for repeat EPDS and diagnostic interview using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) if screen-positive.

Detailed Synthesis

In the study by Ji et al.,[85] results from 534 of 708 enrolled women were analyzed to construct receiver operating characteristic (ROC) curves to determine optimal cutoffs for the four screening instruments used. For diagnosis of postpartum depression, the HRSD-17, HRSD-21, and BDI had an optimal cutoff of 14 in both the early (PP-E) and late (PP-L) postpartum periods (see Table 13). The HRSD-17 was more sensitive but less specific at the PP-E time point compared with the PP-L time point, while the HRSD-21 did not differ in sensitivity by time point. Performance of the BDI was essentially the same at both time points. However, the optimal cutoff for the EPDS increased from 11 at PP-E to 12 at PP-L. The EPDS was more sensitive but less specific at the PP-E time point compared with the PP-L time point.

This study had a high risk of bias. There were multiple screening tests, and the order of administration was not described. The order of administration and the potential for repetitive questions may influence response. Finally, the timing of the followup diagnostic evaluation was not specified.

Table 13. Sensitivity and specificity of screening instruments in the early and late postpartum periods

Instrument and Optimal Cutoff	Time Point	Sensitivity	Specificity
HRSD-17 ≥14	PP-E	81.0%	81.9%
	PP-L	77.4%	84.4%
HRSD-21 ≥14	PP-E	81%	76%
	PP-L	81%	79%
BDI ≥14	PP-E	81.3%	80.3%
	PP-L	82.8%	82.1%
EPDS ≥11	PP-E	83.3%	77.7%
EPDS ≥12	PP-L	76.0%	84.6%

Abbreviations: BDI=Beck Depression Inventory; EPDS=Edinburgh Postnatal Depression Scale; HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression; PP-E=early postpartum period; PP-L=late postpartum period

In the study by Crotty et al.,[101] 625 of 951 women completed a second EPDS at 6 weeks postpartum (66% response rate). Ninety of the 170 women who scored above the cutoff of 12 agreed to diagnostic testing using the SCAN interview. Twenty-three women scoring below 12 also completed the SCAN interview. While the early EPDS identified 58 percent (28 of 48) of women with a confirmed diagnosis of depression, 20 women who subsequently had a high EPDS score at 6 weeks and confirmed depression would have been missed (false negatives). Therefore, while EPDS screening in the immediate postpartum period may identify women at high risk for early development of postpartum depression, sole reliance on screening at this time point is inadequate.

The risk of bias was deemed to be high based on the convenience sampling (recruitment 3 days per week up to a maximum of 10 women per day) and on enrollment up to 3–5 days postpartum, as women who remain hospitalized longer may have more medical complications and thus be at higher risk of postpartum depression.

Key Question 4. Comparative Benefits of Screening

What are the comparative benefits of screening for postpartum depression when compared to no screening, or between different screening strategies (based on choice of screening instrument, timing, setting, etc.)?

Key Points

- Change in a screening instrument depression score was the most common outcome used to assess the comparative benefits of screening for postpartum depression.
- Across a variety of low-intensity interventions, screening was associated with modest improvements in depression.
- One good-quality RCT found that there is benefit to the overall mental health in mothers of screening by health visitors at 6 weeks postnatally.
- The included studies do not allow an assessment of the comparative benefits of screening for postpartum depression by screening instrument, timing, or setting.
- The studies may not be applicable to the general population of women in the United States.
- None of the studies included outcomes for fathers.

Description of Included Studies

Four studies met our inclusion criteria and evaluated the comparative benefits of screening for postpartum depression. Three were RCTs, and one was a quasi-experimental study. In terms of quality, the three RCTs were rated as poor,[102] fair,[103] and good.[104] The quasi-experimental study was rated as poor in quality.[105] The most common relevant outcome was change in a screening instrument depression score. Sample size ranged from 99 recruited at one site[102] to 7,649 enrolled from 101 practices.[104] Only one study was conducted in the United States.[102] This study included only women receiving public assistance and at risk for postpartum depression. The remaining studies were conducted in the UK,[104] Norway,[105] and Hong Kong.[103]

Detailed Synthesis

The earliest of the four studies[102] identified 201 women receiving public assistance and at high risk for postpartum depression based on a predictive risk index.[106] Of these high-risk women, 32 were either already receiving treatment or met criteria for current depression or substance use disorder. Of the remaining 131, 70 (53.4%) were unable to be assessed because of disconnected phones, relocation, or refusal to return calls. The remaining 99 subjects were randomized to either standard antenatal care or to an intervention consisting of four 60-minute group sessions over a 4-week period and a 50-minute individual booster session after delivery. Depressive symptoms were assessed at baseline and 3 months after delivery using the Beck Depression Inventory (BDI). Eighty-six (87%) of the enrollees had followup 3 months after delivery. Two subjects (4%) in the intervention group and eight (20%) in the standard-care group had depression at 3 months based on the depression module of the Longitudinal Interval Follow-up Evaluation. However, there were no differences in depression severity based on the BDI or in social impairment based on the Range of Impaired Functioning Tool. This study was considered to be of poor quality because of the small sample size, differential dropout, and the lack of an intention-to-treat analysis.

A good-quality cluster-randomized trial evaluated screening in 101 general practices in Trent, England.[104] In the intervention practices (n=63), health visitors assessed the mother's mood, administered the EPDS, and provided treatment based on either cognitive-behavioral principles or on person-centered principles. Women were sent a questionnaire by mail at 6 weeks postnatally that included the EPDS and an assessment of other factors potentially associated with depression including social support, stressful life events, threatening experiences, and previous depression. The threshold used for a positive EPDS was 12. Repeat mailings were made to women whose infants were 6, 12, and 18 months old. No specific intervention was given in the control practices (n=38). The EPDS was commonly used in these practices for women whose infants were 6 weeks old. However, unlike the intervention practices, health visitors in in the control group did not provide treatment but instead referred women who screened positive for depression. The SF-12 mental component summary (SF-MCS) and physical component summary (SF-PCS) were used to assess the impact of the intervention at 6 and 12 months. Among those who scored 12 or greater on the EPDS at 6 weeks (n=418), the unadjusted differences in the SF-12 MCS and SF-12 PCS between the control and intervention groups were 4.7 (95% CI, 1.8 to 7.6) and -1.4 (95% CI, -3.5 to 0.7). After adjusting for the 6-week EPDS score, living alone, history of postnatal depression, and any life events, the differences in the SF-12 MCS continued to be statistically significant (5.2 [95% CI, 2.5 to 7.8]) and the difference in the SF-12 PCS remained statistically not significant (-1.7 [95% CI, -3.6 to 0.1]). Across all women (n=2,659) in the trial, differences between the control and intervention groups in SF-12 MCS and SF-12 PCS did not reach statistical significance (p=0.1 and p=0.469, respectively). However, the SF-12 MCS was statistically significantly different in the whole population of women after adjustment for the previously described factors (1.4 [95% CI, 0.5 to 2.3]). Similar differences persisted across the 12- and 18-month assessments. The main potential source of bias for this study was the dropout rate. However, among women in the control group, 87.8 percent had followup at 6 weeks and 74.5% had followup at 6 months; in the intervention group, 82.8 percent had followup at 6 weeks and 68.4% had followup at 6 months. The study was not designed to determine which component of the intervention was effective, but instead looked at whether the entire bundle of interventions could improve outcomes.

A quasi-experimental post-test study[105] compared two areas in Norway, one of which used public health nurses to evaluate women for postpartum depression using the EPDS and clinical assessment and, when necessary, to provide supportive counseling. Screening with the EPDS was offered at 6 weeks and at 3, 6, and 12 months postpartum. The main outcome was the Parenting Stress Index at 12 months postpartum. There was no difference in the overall Parenting Stress Index between the two groups, although there was greater improvement in the EPDS score over time among depressed women in the intervention group (-6.9 in the intervention group versus -4.4 in the control group; p=0.01). However, insufficient data were provided to assess the degree to which this change over time might have been related to the dropout rate. Across all subjects, the dropout rate was 14.5 percent at 3 months, 33.7 percent at 6 months, and 45.8 percent at 12 months. Insufficient data were presented to determine whether the dropout rate differed between groups. This study was rated as poor in quality because of the study design (post-evaluation of a natural experiment) and because of incomplete data about the effect of differential dropout.

A fair quality RCT evaluated the effectiveness of screening for postpartum depression with the EPDS compared with no screening.[103] Participants (n=462) were mothers of 2-month-old babies attending maternal and child health centers in Hong Kong for routine care. In the

intervention group, women attending the centers were screened with the EPDS. Those who scored above the cutoff of 9/10 or answered affirmatively to the suicidal ideation question were then offered nondirective counseling by a maternal and child health (MCH) nurse or by a member of a community psychiatric team (n=231). Women randomized to the control group (n=231) received care as usual, which consisted of clinical assessment by an MCH nurse. If this clinical assessment suggested further management, women in the control group were offered the same services as those in the intervention group, namely, nondirective counseling by an MCH nurse or by a member of a community psychiatric team. The same MCH nurse provided counseling to both groups. Participants in both groups completed a set of questionnaires (including the EPDS) at 6 and 18 months postpartum.

A total of 67 women in the intervention arm screened positive for postpartum depression; of these, 51 received treatment (76.1%). In the control group, 14 women were assessed as having postpartum depression, and 10 received treatment (71.4%). Based on an ITT analysis, fewer women in the intervention group than in the usual care group had EPDS scores above the designed cutoff at 6 months postpartum (13% vs. 22.1%; RR 0.59; 95% CI, 0.39 to 0.89). There were no statistically significant differences between groups on measures of maternal well-being (General Health Questionnaire-12 ([GHQ-12]) or parenting (Parenting Stress Index [PSI]). The study did not assess breastfeeding outcomes. This difference remained statistically significant after adjusting for marital relationship at 2 months, history of psychiatric illness, depression during pregnancy, and relationship with mother-in-law (analysis not reported). However, children of screened women had more visits to the doctors than did children of women in the usual care group (p=0.039). This study was considered to be of fair quality because there appear to be baseline differences in the groups with lack of clarity on how adjustment was performed; these differences also suggest potential problems with the randomization process. In addition, missing data were imputed by means of group substitution at followup, which may bias the findings.

These four studies do not allow an assessment of the comparative benefits of screening for postpartum depression by screening instrument, timing, or setting. However, the good-quality RCT[104] suggests that screening provides a benefit to overall mental health in mothers based on the SF-12 MCS. It is unclear how the observed benefit translates into improved quality of life, family functioning, or health outcomes. Interpreting changes in depression screening scores is challenging because of the fluctuations in these scores over time.

Key Question 5. Comparative Harms of Screening

What are the comparative harms of screening for postpartum depression when compared to no screening, or between different screening strategies (based on choice of screening instrument, timing, setting, etc.)?

Key Points

- Only one study reported potential harms of screening for PPD. Children of women randomized to screening had more doctor visits, with no difference in baseline health, than did women in a control group.

Description of Included Study

Only one study met the inclusion criteria for KQ 5; this was a fair-quality RCT.[103] Most women in the study were married (95.5%) and had no past history of psychiatric illnesses (98.4%).

Detailed Synthesis

In the Hong Kong RCT described above,[103] children of screened women had more doctor visits (mean 2.39; 95% CI, 2.07 to 2.7)) compared children of women in the usual care group (mean 1.97; 95% CI, 1.72 to 2.21;, p=0.039) at 3 months, without any evidence of differences in child health status. This difference was no longer significant at 18 months (mean visits in screened group 5.14; 95% CI, 4.57 to 5.71 mean visits in control group 4.97; 95% CI, 4.58 to 5.36). This study was considered to be of fair quality because there appear to be baseline differences in the groups with lack of adjustment. In addition, missing data were imputed by means of group substitution at followup, which may bias the findings.

Key Question 6. Factors Affecting the Likelihood of an Appropriate Action After a Positive Screening Result

Is the likelihood of an appropriate action (referral, diagnosis, treatment, etc.) after a positive screening result affected by timing, setting, patient characteristics, or other factors?

Key Points

- The EPDS was the most common screening tool used across studies.
- Overall rates of referral and treatment were low, ranging from 5–30 percent, for women who screened positive for postpartum depression.
- One good-quality cross-sectional study found that women who screened positive for postpartum depression at delivery had a higher proportion of psychiatric followup than those who screened positive prenatally or at 6 weeks postpartum.
- A fair-quality prospective cohort study reported that women who screened positive at 6 weeks postpartum had lower rates of referral and treatment for symptoms of anxiety and depression than women who screened positive during the third trimester.

Description of Included Studies

Five studies met the inclusion criteria for KQ 6. Two were prospective cohort studies,[107,108] one was a cross-sectional study,[109] one was a pre-post intervention study,[110] and one was a quasi-experimental design.[111] The two cohort studies were rated as fair quality,[107,108] the cross-sectional study was rated as good quality,[109] while the pre-post intervention study[110] and quasi-experimental study[111] were rated as poor quality. Four studies were conducted in the United States[107,109-111] and one in Germany.[108] All five provided some measure of appropriate diagnosis and treatment of depression. Screening most commonly occurred in the first 8 weeks postpartum; four of the five studies used the EPDS as the screening tool.

Detailed Synthesis

A good quality cross-sectional study[109] assessed 293 U.S. women at either 36 weeks gestation, delivery, or the 6-week postpartum visit with the self-completed EPDS. The stated goal of the study was to assess the most advantageous timing for postpartum depression screening that optimized access to care. A cutoff of 10 was used to signal probable postpartum depression, and if a woman screened positive she was offered followup psychiatric services. The study assessed rates of psychiatric followup care for all women who screened positive at each time point. Overall, 12.6 percent of women screened positive for postpartum depression. However, prevalence varied across time: 5 percent screened positive at 36 weeks, 16 percent screened positive at delivery, and 14 percent screened positive at 6 weeks postpartum. Among those with positive screens, the proportion receiving psychiatric evaluation varied significantly with timing: 33 percent completed evaluations at 36 weeks, 100 percent completed evaluations at delivery, and 15 percent completed evaluations at 6 weeks postpartum ($p<0.001$). Prenatal and 6-week postpartum evaluation took place in outpatient settings, while the delivery assessment took place prior to discharge from the hospital, which likely contributed to the rates of completed evaluations. Of the 37 women who screened positive, 20 (54%) were subsequently diagnosed with depression, and 19 percent of these started treatment for depression.

A fair-quality prospective cohort study followed 772 German mothers screened for postpartum depression.[108] First, women were screened with the EPDS at 6–8 weeks postpartum. Those mothers scoring 9.5 or higher were offered repeat screening with the EPDS 3 weeks later. If women scored a 9.5 or higher on rescreening, they were assessed further with the HAM-D and a DSM-IV classification by clinical judgment by a clinical psychologist. Treatment was offered to mothers with a diagnosis of depression as assessed by DSM-IV criteria. At the first EPDS assessment at 6–8 weeks postpartum, 17 percent (n=132) of mothers screened positive for postpartum depression. These women were then rescreened with the EPDS 3 weeks later; 3.6 percent (n=28) of the women with an elevated EPDS on the initial screen tested positive for postpartum depression on rescreening, and all met DSM-IV criteria for depression. Of the 28 women diagnosed with postpartum depression, only 5 (18 percent) accepted treatment. No multivariate analysis was performed to assess predictors of referral, diagnosis, or treatment for depression.

Another fair-quality prospective cohort study sought to examine detection, treatment, and referral of both postpartum depression and anxiety by obstetrical providers during pregnancy and at 6 weeks postpartum among 491 U.S. women.[107] Postpartum depression was assessed with the EPDS, and a cutoff of 10 was used to indicate a positive screening result. Anxiety was assessed using the anxiety portions of the Patient Health Questionnaire (PHQ). Obstetrical medical record reviews were used to assess documentation of mental health diagnosis, referral and treatment. A total of 22.2 percent of women screened positive for postpartum depression and 4.3 percent were positive for an anxiety disorder during the prenatal assessment in the third trimester. Only 46 of 113 women (41%) who screened positive during the third trimester had documentation of psychiatric symptoms or diagnosis by a provider in the medical record. Of those with medical records documentation, only 37 percent had further documentation of mental health treatment, and 43 percent (n=20) had documentation of a referral. Only 10 of the referred women (50%) accessed the referral. Thus, only 15 percent of women who screened positive for postpartum depression or anxiety had documentation of treatment during pregnancy, and an additional 18 percent had documentation of a referral for treatment. At 6 weeks postpartum, 17 percent (51 of 299) screened positive for postpartum depression and anxiety. Of this 17 percent, only 29.4

percent had documentation of psychiatric symptoms or diagnosis in the medical records, but nearly all (93%) had subsequent documentation of treatment or referrals for mental health as assessed by medical record review. Overall, only 27.5 percent (14 of 51) of women who screened positive for anxiety or depression at 6 weeks postpartum received any treatment or referral for mental health services. Thus, documented rates of referral and treatment were low overall. Women who screened positive at 6 weeks postpartum had slightly lower rates of referral and treatment for symptoms of anxiety and depression compared with women who screened positive during the third trimester (27.5% vs. 33%; p value NR). Again, no multivariate analysis was performed to assess predictors of referral, diagnosis, or treatment for depression.

A poor-quality quasi-experimental study[111] sought to examine the program impact of a Healthy Start depression treatment initiative in New Haven, Connecticut. The Healthy Start depression initiative consisted of mental health assessment with the PRIME-MD and Brief Patient Health Questionnaire (BHQ) and referrals to services. Women who had depression could also attend weekly drop-in services that provided behavioral and pharmacological treatment. The study constructed three cohorts to assess the impact of Healthy Start on depression detection, referral, and treatment: a pre-Healthy Start depression initiative cohort, a post-Healthy Start cohort that was enrolled in the depression initiative, and a post-Healthy Start cohort that was not enrolled in the depression initiative. Propensity scoring was used to control for imbalance of baseline covariates. Rates of depression detection (p=0.003) and referral (0<0.001) were significantly different among the three groups, with the pre-Healthy Start group demonstrating the highest rates. The proportion of women in treatment for depression was not significantly different across groups (p=0.077); only 0.3 percent of women in the pre-Healthy Start group, 2 percent of women in enrolled Healthy Start, and 1 percent of women not enrolled in the Healthy Start program were in treatment for depression. As this was a quasi-experimental study, it suffered from numerous issues with selection, detection, and performance biases contributing to its poor quality rating.

A poor-quality pre-/post-intervention study evaluated a brief obstetric clinic-based intervention on perinatal depression treatment in the context of a newly implemented policy of routine screening at a university-affiliated obstetric clinic in the United States.[110] In accordance with the new policy, all women were screened at their first prenatal visit with the EPDS, and a score of 10 was used to signal probable depression and to prompt referral for further evaluation and treatment. A total of 1,298 new obstetric patients were screened for depression in accordance with this policy from November 2002 to January 2004. A total of 207 women (16%) scored above 10 on the EPDS, and 73 of these (35%) consented to be in the study and completed baseline interviews, which occurred 2 weeks after the second prenatal visit. The baseline survey included the Mood Disorders Module of the SCID for DSM-IV to obtain diagnosis of current or past depression. Depression treatment was assessed by self-report. Women were interviewed again 1 month after baseline and 6 weeks postpartum. The SCID was repeated at the 6-week postpartum interviews. The intervention consisted of notification to the treating physician of an elevated EPDS score via a flag in the medical record and nurse-delivered feedback to the patient on depression score, education about depression, and a referral for the patient occurring before the second prenatal visit. Based on medical record review and study interviews, authors constructed four time points for the assessment of depression treatment: Time 1, 3 months prior to the first prenatal visit; Time 2, time between first prenatal care visit and baseline prenatal interview; Time 3, time between the baseline prenatal interview and 1-month prenatal interview; and Time 4, time between 1-month prenatal interview and 6-week postpartum interview.

At baseline, 40 percent of the women in the study who screened positive for postpartum depression with the EPDS met diagnostic criteria for depression. At Time 1(3 months prior to first prenatal visit), 16 percent of women with an EPDS of 10 or more were receiving some form of depression-related treatment as assessed by medical records review. At Time 2 (after EPDS screening and intervention), 21 percent of EPDS screened positive women self-reported that they were receiving treatment for depression, and this proportion remained constant through Time 3 (one month after baseline interviews). By 6 weeks postpartum (Time 4), 18 percent of EPDS screened positive women reported receiving treatment for depression. As part of the baseline survey (which occurred after routine EPDS screening and second prenatal visit), women were also asked if their physicians had discussed their elevated EPDS scores with them. The majority (67%) reported that their physician did discuss depression during their prenatal visit. Assessing a limited number of covariates, study investigators modeled the likelihood of depression treatment throughout the study using multivariable logistic regression. The only significant predictors of depression treatment were treatment prior to EPDS screening and greater depression severity as measured by the BDI-II. This study had a small, highly selected sample of women; only 35 percent (72 of 207) of women with an elevated EPDS consented to participate in the study and were followed over time.

Discussion

Key Findings and Strength of Evidence

In this comparative effectiveness review (CER), we reviewed 36 unique studies represented by 40 publications that evaluated tools for screening for postpartum depression, risk factors for postpartum depression, and factors influencing the effectiveness of screening for postpartum depression. The available evidence did not allow us to draw any conclusions about the balance of benefits and harms of screening for postpartum depression, or whether specific tools or strategies would result in a more favorable balance.

KQ 1. Performance Characteristics of Screening Instruments

Although the included studies varied widely in country, language, setting, and timing of testing, estimates for both sensitivity and specificity were in the 80–90 percent range for most of the screening tests for which there was evidence. As expected, there was an inverse correlation between sensitivity and specificity. Increased sensitivity was associated with decreased specificity when the threshold for an abnormal screening test was varied both within and between studies.

Multiple studies were available only for the Edinburgh Postnatal Depression Scale (EPDS) and the Postpartum Depression Screening Scale (PDSS). Although heterogeneity in both the clinical characteristics of the population being screened and the threshold used precluded quantitative synthesis, the range of observed sensitivity and specificity for both of these tests fell within the 80–90 percent range. In the two studies that directly compared these two instruments, confidence intervals (CIs) for both sensitivity and specificity overlapped.

One Hungarian study of the 24-item Leverton Questionnaire reported sensitivity of 95.2 percent (95% CI, 90.4 to 98.1%) and specificity of 91.3 percent (95% CI, 88.4 to 93.7%). We did not identify any confirmatory studies in a U.S. setting.

Table 14 summarizes the strength of evidence for each screening test reviewed.

Table 14. Strength of evidence domains for test characteristics of screening tests for postpartum depression

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Antenatal Risk Questionnaire						
Sensitivity	1 (276)	High	NA	Direct	Imprecise	Low SOE 78.1% (65.0 to 88.7%)
Specificity	1 (276)	High	NA	Direct	Imprecise	Low SOE 47.1% (40.3 to 59.9%)
BDI						
Sensitivity	2 (1,475)	Medium	Consistent	Direct	Imprecise	Low SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
Specificity	2 (1,475)	Medium	Consistent	Direct	Precise	Low SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
EPDS						
Sensitivity	10 (3,847)	Medium	Consistent	Direct	Imprecise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
Specificity	10 (3,847)	Medium	Consistent	Direct	Precise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
HSRD-17						
Sensitivity	1 (1376)	High	NA	Direct	Imprecise	Low SOE 80–85% (Range of point estimates across thresholds)
Specificity	1 (1,376)	High	NA	Direct	Imprecise	Low SOE 80–85% (Range of point estimates across thresholds)

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
HSRD-21						
Sensitivity	1 (1,375)	High	NA	Direct	Imprecise	Low SOE 80–85% (Range of point estimates across thresholds)
Specificity	1 (1,375)	High	NA	Direct	Imprecise	Low SOE 75–80% (Range of point estimates across thresholds)
Leverton Questionnaire						
Sensitivity	1 (617)	Low	NA	Direct	Imprecise	Low SOE 95.2% (90.4 to 98.1%)
Specificity	1 (617)	Low	NA	Direct	Imprecise	Low SOE 91.3% (88.4 to 93.7%)
PDSS						
Sensitivity	4 (904)	Medium	Consistent	Direct	Imprecise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
Specificity	4 (904)	Medium	Consistent	Direct	Precise	Moderate SOE 80–90% (Approximate range of point estimates at most commonly used thresholds)
PHQ						
Sensitivity	1 (506)	Low	NA	Direct	Imprecise	Low SOE 75–89% (Range of point estimates at varying threshold—wide 95% CIs for point estimates at each threshold)
Specificity	1 (506)	Low	NA	Direct	Imprecise	Low SOE 83–91% (Range of point estimates at varying threshold)
Two-Question Screen						
Sensitivity	1	Low	NA	Direct	Imprecise	Low SOE 100% (93.3 to 100%)
Specificity	1	Low	NA	Direct	Imprecise	Low SOE 44.3% (39.5 to 48.6%)

Abbreviations: BDI=Beck Depression Inventory; CI=confidence interval; EPDS=Edinburgh Postnatal Depression Scale HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression; NA=not applicable; PDSS=Postpartum Depression Screening Scale; PHQ=Patient Health Questionnaire

The probability of a false negative or false positive test result is a function of test sensitivity, specificity, and the prevalence of the underlying disorder. Table 15 illustrates the interaction of these three parameters, using the 80–90 percent range for sensitivity and specificity observed for most of the studies in our review. The estimated point prevalence of major depression at various points in the first 12 months after delivery in the 2005 Agency for Healthcare Research and Quality (AHRQ) Report was in the 4–8 percent range;[3,4] the prevalence in the majority of studies included in this review was in the 10–20 percent range. As seen in Table 15, the number of false positives well exceeds the number of true positives at lower prevalence, especially when prevalence is below 10 percent.

Table 15. Effect of prevalence of major depression on true positives, false positives, and false negatives at varying levels of sensitivity and specificity

Prevalence	Sensitivity 90% Specificity 80%	Sensitivity 85% Specificity 85%	Sensitivity 80% Specificity 90%
4%			
True Positives	3.60%	3.40%	3.20%
False Positives	19.20%	14.40%	9.60%
False Negatives	0.40%	0.60%	0.80%
8%			
True Positives	7.20%	6.80%	6.40%
False Positives	18.40%	13.80%	9.20%
False Negatives	0.80%	1.20%	1.60%
15%			
True Positives	13.50%	12.75%	12.00%
False Positives	17.00%	12.75%	8.50%
False Negatives	1.50%	2.25%	3.00%

We did not identify any studies that compared the ability of individual items in specific instruments to correctly identify particular signs or symptoms of depression. One study found moderate agreement between the suicidal ideation item of the EPDS and a diagnostic instrument, but suicidal ideation was not significantly associated with any outcomes, including response to therapy.

KQ 2. Effect of Individual Subject Factors on Screening Performance

Table 16 summarizes the strength of evidence for the individual factors identified in the included studies. Women with a history of previous psychiatric disorders, particularly mood disorders, and women in a poor-quality relationship or with low levels of social support, are at higher risk for postpartum depression. Although the heterogeneity in populations and instruments used to measure these domains precluded quantitative synthesis, the results were consistent across studies, with relatively large odds ratios of 2.0 or more, and were almost always statistically significant in multivariate analyses. Although strength of evidence for some individual risk factors within these broad categories was low (primarily based on single studies or wide CIs), the overall consistency leads to an assessment of moderate strength of evidence.

Chronic medical conditions and adverse pregnancy outcomes were also consistently associated with postpartum depression, but the smaller number of studies assessing these factors led to a low strength of evidence rating.

With the exception of unemployment, there was insufficient evidence to assess the association between other maternal demographic factors and postpartum depression.

We did not identify any studies meeting our inclusion criteria which evaluated a risk prediction instrument (similar to the Gail model for predicting breast cancer risk[112]).

Table 16. Strength of evidence domains for associations with patient characteristics and risk of postpartum depression

Risk Factor	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Maternal demographics						
Age	2 (4,773)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Education	2 (4,773)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Income	2 (4,998)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Employment status (unemployed vs employed)	1 (363)	High	NA	Direct	Imprecise	Low SOE OR 2.8 (1.1. to 4.9)
Obstetric history						
Parity	2 (4,998)	Medium	Consistent	Direct	Imprecise	Insufficient
Preterm/low birthweight infant	2 (4,711)	Medium	Consistent	Direct	Precise	Low SOE for Increased risk of postpartum depression
Smoking	2 (4,773)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Alcohol use	1 (4,348)	Medium	NA	Direct	Imprecise	Insufficient
General medical history						
Poor health status/chronic illness	2 (4,773)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Obesity	1	Medium	NA	Direct	Imprecise	Insufficient
Psychiatric history						
History of perinatal depression	2 (640)	High	NA	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
History of depression	5 (1,440)	Medium	Consistent	Direct	Precise	Moderate SOE for Increased risk of postpartum depression
History of premenstrual dysphoric disorder	1	Medium	NA	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Any psychiatric diagnosis	2 (1,025)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Anxiety	1	Medium	NA	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Personality (vulnerable/ neuroticsm)	2 (777)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression

Risk Factor	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Relationship/social support						
Marital status (single/no relationship)	2 (4,773)	Medium	Consistent	Direct	Imprecise	Low SOE for Increased risk of postpartum depression
Poor relationship quality	5 (5,856)	Medium	Consistent	Direct	Imprecise	Moderate SOE for Increased risk of postpartum depression
Poor social support	4 (1,333)	Medium	Consistent	Direct	Imprecise	Moderate SOE for Increased risk of postpartum depression

Abbreviations: CI=confidence interval; NA=not applicable; OR=odds ratio; SOE=strength of evidence

KQ 3. Effect of Testing Variables on Screening Performance

We identified only two studies that provided estimates of test performance based on timing, and the evidence was insufficient to assess whether the timing of screening relative to delivery affects sensitivity or specificity for any screening instrument. In one study judged to be at high risk of bias, test characteristics for four different screening instruments were similar when measured in the first 8 weeks after delivery compared with 2–6 months after delivery. We did not identify any studies comparing screening instrument performance across settings or type of provider (Table 17).

Table 17. Strength of evidence domains for the effect of varying timing, setting, and provider on screening for postpartum depression

Variable	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Timing						
Delivery to 8 weeks vs 8 weeks to 6 months	1 (534)	High	NA	Direct	Imprecise	Insufficient
Delivery vs 6 weeks	1 (625)	High	NA	Direct	Imprecise	Insufficient

Abbreviations: CI=confidence interval; NA=not applicable; SOE=strength of evidence

KQ 4. Comparative Benefits of Screening; and KQ 5. Comparative Harms of Screening

We identified some evidence of benefit to screening compared with no screening or usual care, either through identifying higher risk women prior to delivery and implementing primary preventive strategies, or through screening and referral for treatment. Screening led to decreases in depressive symptoms as measured by repeated administration of the screening instruments themselves (low strength of evidence due to poor- to fair-quality studies and imprecise estimates), improvement in the mental health component of a health-related quality-of-life

instrument (low strength of evidence primarily due to a single fairly small study), and possibly parental stress (low strength of evidence due to inconsistent results across studies; Table 18).

One fair-quality study found a statistically significant increase in the number of unscheduled doctor visits in the first 3 months after delivery for infants of screened women compared with unscreened women, but this difference was no longer significant by 12 months. None of the other studies addressed potential harms of screening.

We did not identify any evidence that choice of screening instrument, timing of screening, setting, provider, or other factor affected the outcomes of screening.

Table 18. Strength of evidence domains for benefits and harms of screening for postpartum depression

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Benefits						
Depressive symptoms	3 (1171)	Medium	Inconsistent	Direct	Imprecise	Low SOE for Reduced number of symptoms with screening and intervention
Mental health score (SF-12)	1 (418)	Low	NA	Direct	Imprecise	Low SOE for Improved scores with screening and intervention
Parental stress	2 (853)	Medium	Inconsistent	Direct	Imprecise	Insufficient
Harms						
Unscheduled doctor visits for infant	1 (430)	Medium	NA	Direct	Imprecise	Low SOE for Increased number of visits for infants of screened women

Abbreviations: CI=confidence interval; NA=not applicable; SF-12=Medical Outcomes Study 12-Item Short-Form Health Survey; SOE=strength of evidence

KQ 6. Factors Affecting the Likelihood of an Appropriate Action After a Positive Screening Result

In general, rates of followup in women with positive screening test results in all of the studies included across all Key Questions (KQs) were low, ranging from 5–50 percent. Differences in country, setting, population characteristics, screening instrument, and timing precluded synthesis across studies. Only two studies allowed direct comparison of rates at different times during pregnancy and the postpartum period (Table 19): one study found significantly higher rates of referral when screening was performed during the delivery admission (100%) compared with 36 weeks gestation (33%) or at 6 weeks postpartum (15%; $p<0.001$),[109] while a second found a much smaller difference when comparing prenatal (33%) with postpartum (27%) screening (p=not statistically significant [NS]).[107] We did not identify any studies that addressed potential differences based on setting or provider.

Table 19. Strength of evidence domains for the effect of timing, setting, and provider on rates of referral and treatment among women with a positive screening test for postpartum depression

Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Timing						
Prenatal vs postpartum	2 (784)	Medium	Consistent	Direct	Imprecise	Low SOE for Higher rates of referral/diagnosis prenatally
Delivery vs postpartum	1 (293)	Low	NA	Direct	Imprecise	Low SOE for Higher rates of referral/diagnosis during delivery admission

Abbreviations: CI=confidence interval; NA=not applicable; SOE=strength of evidence

Findings in Relationship to What is Already Known

Our review focused on studies published subsequent to the 2005 AHRQ Evidence Report on perinatal depression.[3,4] Key findings of the AHRQ review included:

- Patient characteristics in studies which did not reflect the diversity of the U.S. population of pregnant and postpartum women;
- Lack of precision for estimates of test characteristics, particularly for test sensitivity;
- Widely overlapping CIs for estimates, precluding indirect comparison across tests;
- Relatively few studies directly comparing results of multiple screening instruments;
- Overall better screening instrument sensitivity for major depression compared to major and minor depression combined; and
- No studies comparing screening with no screening.

Recommendations included:

- Designing and powering studies to improve the precision of sensitivity estimates, if a premium is placed on negative predictive value of screening;
- Including more diverse populations in studies;
- Direct comparison of different screening instruments within studies;
- Studies that evaluate a broader range of timing;
- Designing studies comparing screening with no screening.

Our findings were broadly consistent with these results. We did identify some studies that included more diverse U.S. populations (including the development of a Spanish-language version of one of the instruments for Latina populations[91]), studies directly comparing different screening instruments,[76,77,85] and studies comparing screening with no screening.[103-105] However, the overall strength of the evidence base is not much better now than it was in 2005, although given the amount of time needed to design, implement, analyze, and report trials of the size necessary to address many of these concerns, it is likely that most if not all studies that considered the 2005 report in their design have not yet been published (note that we are aware of one in-press paper which is also a comparison of screening vs. no screening, which will hopefully be available prior to submission of the final version of this report).

Applicability

The effects of interventions, as determined in research studies, do not always translate well to usual practice, where patient characteristics, clinical training, diagnostic workup, and resources may differ importantly from study conditions. Thus, we assessed the applicability of the included studies.[113]

Many included studies recruited populations where large differences between demographics of study population and community patients existed. Overall, only about one-third of included studies were conducted in the United States; the majority were conducted in Europe or the UK (48%). Event rates for postpartum depression between countries differ significantly due to dissimilarities in social and cultural contexts (e.g., family structures, gender roles). Moreover, the health care system in the United States differs considerably from those in Europe and the UK, making it problematic to translate findings to the U.S. context. Many studies had highly selected samples due to high rates of non-response or attrition during the studies, which limits these findings to broader populations. The majority of studies were conducted in women in their late twenties to early thirties. Few studies were conducted with samples of older maternal age. Finally, the prevalence of major depression in studies estimating the sensitivity and specificity was substantially higher than U.S.-population-based point-prevalence estimates, suggesting that that the positive predictive value of any screening instrument in a low-risk population will be substantially lower than the estimates derived from validation studies.

The EPDS is the most widely known and used screening tool for postpartum depression: over two thirds of studies assessed postpartum depression with the EPDS. To the extent that the EPDS is considered "standard of care," findings from these studies would have reasonable applicability. However, these studies used a range of cutoffs to signal probable postpartum depression (range: 8–13), and descriptions of testing protocols were not specific enough to inform routine clinical care. Confidence intervals for sensitivity estimates for all screening tests were wide, and for the most part sensitivity and specificity estimates were qualitatively similar. In addition, some studies administered the screening test in the perinatal through discharge period in a hospital setting—the results from this setting may not be representative of the results for screening in outpatient settings.

There were few direct comparisons between screening instruments, and the studies that did directly compare instruments did not identify substantial differences. There were only a few studies that directly compared screening with any instrument with no screening, and, although they suggest an improvement in depressive symptoms, there is limited data on other maternal or infant health outcomes. Lastly, there is limited information on paternal outcomes.

Implications for Clinical and Policy Decisionmaking

The 2005 AHRQ report concluded that there was a lack of evidence on the overall effectiveness of screening for depression in pregnancy or the postpartum period, lack of consensus on the appropriate target for screening (major depression alone vs. major and minor depression), and, if screening is performed, uncertainty about which instrument to use. These uncertainties are reflected in the recommendations by various stakeholder organizations discussed in the Introduction, and the evidence reviewed for this report does little to resolve those uncertainties: we found some evidence that screening improves some maternal outcomes compared with no screening, but the overall effect of this improvement on longer term maternal and infant outcomes is unclear.

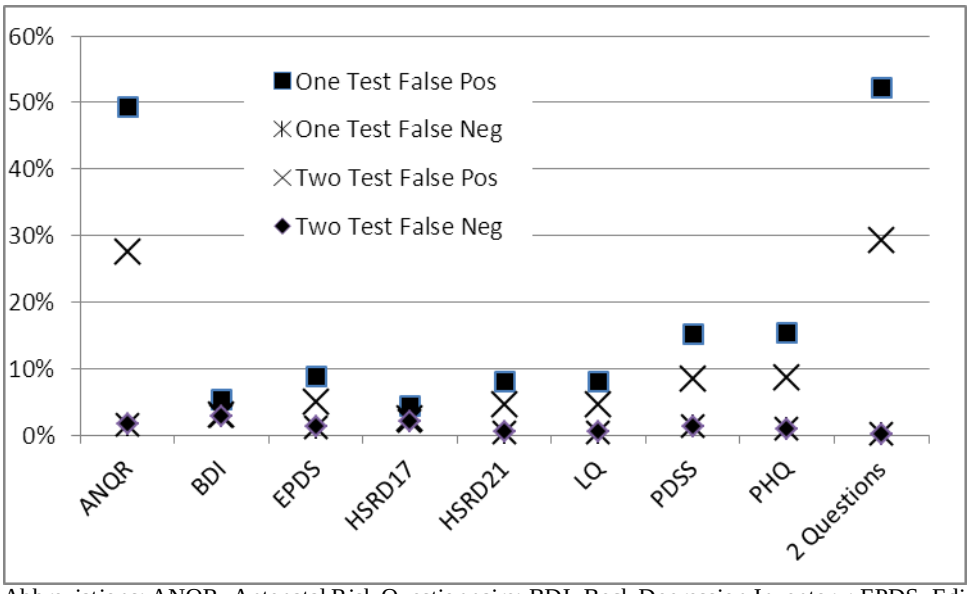
The issue of the number of signs and symptoms of depression to include as the diagnostic threshold for screening is an important one. There are effective treatments for major depression, but there is less evidence of effectiveness for treatment aimed at specific symptoms, or for clusters of symptoms that do not meet the threshold for a definition of major depression. The rationale for screening is stronger if the cumulative prevalence of treatable symptoms is high, even if the joint probability of enough symptoms to meet diagnostic criteria for major depression (i.e., the prevalence of "major depression") is low.

Potential Value of Simulation Modeling

The lack of evidence for the benefits and harms of screening ultimately contributes to the difficulty in identifying the optimal screening test and strategy. There is clearly a trade-off between false positive and false negative test results (Table 15). Given estimates of the point prevalence of depression of 3–7 percent in the postpartum period[3,4] and the range of sensitivities and specificities of the most commonly used screening instruments, it seems likely that the number of false positive results are likely to exceed the number of true positive results with the use of any single screening instrument. In the absence of direct evidence, one method for estimating the balance of benefits and harms is to use a simulation model. As described in the Methods, we adapted an existing model of pregnancy, the postpartum period, and infancy[69] to generate preliminary estimates of these trade-offs using the available evidence, including the existing uncertainty surrounding the estimates of sensitivity and specificity for currently available tests.

One strategy to reduce the number of false positive results would be to use serial testing with a highly sensitive test first, followed by a highly specific test in patients with positive results on the first test. One possible option would be to use the two-question screen, which had a reported sensitivity of 100 percent and specificity of 44 percent, followed by a second screening test in women with a positive answer to either of the two questions, as suggested by Gjerdingen et al.[81] Figure 9 shows the expected proportions of false positives and false negatives for each of the nine screening tests alone and as a followup in patients who give a positive response to one of the two questions, assuming a prevalence of postpartum depression of 5.8 percent at 2 months postpartum (the highest point prevalence estimate in the 2005 AHRQ report), using randomly selected point estimates from the studies reviewed for KQ 1. Serial testing has no effect on false negative rates, but substantially decreases false positive rates for all tests, with the decrease being most dramatic for tests with lower specificity (for clarity, error bars for the estimates are not shown, but overlap—the figure should not be used to draw inferences for between-test comparisons)

Figure 9. Estimated false negatives and false positives for each of nine tests as either a single screen ("One Test") or as part of a serial testing strategy following a positive response to one of the two questions ("Two Test") at a prevalence of 5.8%

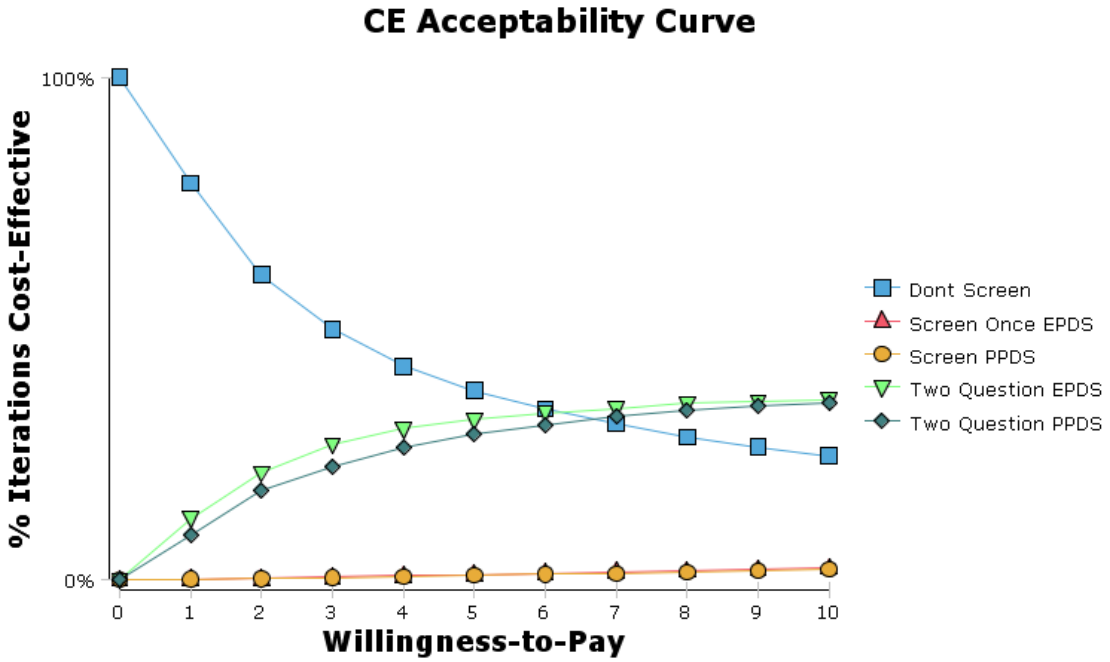


Abbreviations: ANQR=Antenatal Risk Questionnaire; BDI=Beck Depression Inventory; EPDS=Edinburgh Postnatal Depression Scale HRSD-17=17-Item Hamilton Rating Scale for Depression; HRSD-21=21-Item Hamilton Rating Scale for Depression; LQ=Leverton Questionnaire; PDSS=Postpartum Depression Screening Scale; PHQ=Patient Health Questionnaire

Better understanding of the trade-offs between harms and benefits would also help with identifying the optimal test and strategy. As an example, Figure 10 presents the results of a microsimulation comparing no screening, screening with the EPDS alone, screening with the Postpartum Depression Screening Scale (PDSS) alone, screeing with two questions followed by the EPDS, or screening with two questions followed by the PDSS. For each simulation (n=10,000), the value for test sensitivity and specificty were randomly drawn from the distributions described in each study described in KQ 1 (where the probability of a specific study being chosen was a uniform distribution, the specificity was drawn from a beta distribution based on the study-specific values, and the sensitivity was drawn from a function based on the selected specificity value and a log-normal distribution of the study-specific diagnostic odds ratio, in order to account for the negative correlation between sensitivity and specificity[114]). Prevalence was drawn from a beta distribution based on the estimated point prevalence at 2 months in the 2005 AHRQ report. Results are shown as an "acceptability curve," where the trade-off between false positives (equivalent to costs in a cost-effectiveness analysis) and treated depression (the measure of effectiveness) is considered using a "willingness-to-pay" threshold—in this case, how many false positives per treated depression is a decisionmaker willing to accept? The optimal strategy is the one that has the highest net value at a given willingness-to-pay. The x-axis varies the ratio of false positives to detected cases from 0 to 10, while the y-axis depicts the proprortion of simulations where a given strategy was optimal. For example, if no false positives are acceptable, then no screening is always optimal, given that none of the screening strategies has a specificity of 100 percent. As the "acceptable" ratio increases, the proportion of simulations where no strategy would be preferred to any of the alternatives decreases. Values of acceptabilty where there is little difference between strategies indicate that the uncertainty surrounding the values of the parameters is too great to distinguish between them.

Figure 10 shows that serial testing is almost always favored over a single test; that there is minimal difference between the EPDS and PDSS given the available evidence; and that, even with serial testing, there is likely to be a high number of false positives associated with screening. With additional evidence on the clinical harms (as well as costs) associated with a false positive result, making a recommendation for or against screening, either generically or with a specific test, would be much easier.

Figure 10. Acceptabilty curve for trade-off between false postives ("costs") and treated depression ("effectiveness") at different thresholds for false positives/treated depression ratio ("willingness-to-pay"). The y-axis illustrates the proportion of simulations (n=10,000) where a given strategy was optimal at a given acceptabilty threshold for the ratio of false positives/treated depression.



Abbreviations: EPDS=Edinburgh Postnatal Depression Scale; PDSS=Postpartum Depression Screening Scale; [To be added]

Consensus on the relative importance of false positives and false negatives will also help in selecting study thresholds, or in the design of new screening strategies. Many of the studies we reviewed selected a screening threshold based on the value that maximized the area under the receiver operating characteristic (ROC) curve. If a false positive and a false negative are equally bad, then choosing the threshold that optimizes both is reasonable; however, if the relative importance of the outcomes associated with each incorrect test result is different, then that difference needs to be included in the criteria for selecting the threshold. The frequency of testing, along with the natural history of the target condition, is also important—if the target condition is unlikely to worsen between screening intervals, then optimizing specificity over sensitivity might be reasonable, whereas optimizing sensitivity might be better for a one-time screen.

Followup rates for women with positive screening results were uniformly low in the studies reviewed. The impact of these low followup rates on the overall effectiveness of screening is unclear. On the one hand, given the high false positive rate of most of the screening instruments studied, if the majority of women who did not get further evaluation represented women who

were truly not depressed, then the overall effectiveness of screening might not be substantially worsened. On the other hand, to the extent that women with true positives results are equally likely (or even more likely) to not follow up as women with false positive results, screening effectiveness (and cost-effectiveness) is adversely affected. Without either better evidence about the possibility of differential followup rates, or systems in place to maximize appropriate followup for screen positives, implementing screening could lead to a significant waste of resources, including both provider and patient time.

Although we did not find evidence for substantial differences in screening instrument performance based on timing relative to delivery, there was some evidence for higher rates of followup when screening was performed closer to delivery. Given that the risk for postpartum depression appears to continue at least through the first 12 months after delivery, the optimal timing for screening relative to both test performance and followup is unclear. The ongoing risk suggests that screening throughout the postpartum period might be optimal for maximizing the detection of depression, but, the more frequent the screening, the greater the number of false positives, as has clearly been demonstrated with cancer screening models.[115]

Limitations of the Comparative Effectiveness Review Process

There were several limitations to our review. We limited our search to English-language articles for two main reasons: a lack of translation resources, and a priority for studies that were applicable to U.S. populations. It was the opinion of the investigators and the Technical Expert Panel (TEP) that the resources required to translate non-English articles would not be justified by the low potential likelihood of identifying relevant data unavailable from English-language sources. To the extent that studies relevant to screening for postpartum depression in the U.S. population might be published in languages other than English, we may have failed to include relevant studies.

Because there was substantial overlap between our KQs and the KQs considered in the 2005 AHRQ review, we focused our search on articles published subsequent to the last date in the search conducted for that report. The major overlap in topic between the two reports in the test characteristics of specific screening instruments; it is possible that abstraction of some of the articles included in the 2005 report might have allowed formal synthesis of sensitivity/specificity estimates for some tests at some thresholds; however, given the heterogeneity between studies, it seems unlikely that any additional clarity about relative test performance would have been achieved.

We restricted included articles on test performance and outcome to those which used a reference diagnostic interview or instrument in all positive subjects and all or a random sample of screen negatives. The low rates of followup for clinical diagnosis are also seen in research studies, which may lead to selection bias in studies which require a reference standard.[116] To the extent that the effective interventions are available for specific symptoms detected by a screening instrument, even if diagnostic criteria for depression are not met, this requirement may also underestimate some of the clinical benefits of screening.

Limitations of the Evidence Base

As noted above, many of the limitations of the evidence base noted in the 2005 AHRQ report[3,4] are still present. These include

- Patient characteristics in studies which do not reflect the diversity of the U.S. population of pregnant and postpartum women, or which are focused on high-risk populations only. Although we identified some studies in more diverse populations, additional studies are needed, particularly given the need to increase the precision of estimates of test characteristics and the potential for variations in prevalence of depression across populations.
- Relatively few high-quality studies comparing results for multiple screening instruments, either through randomization or by administering different instruments to the same subject.
- Relatively few high-quality studies comparing screening to no screening; we identified only one fair-quality randomized controlled trial (RCT), although we are aware of at least one additional RCT that is in press.
- Lack of evidence for any benefit associated with detecting symptoms of depression that together do not meet criteria for a diagnosis of major depression. Such evidence would be extremely helpful in setting thresholds for a positive test, as well as helping define the overall benefits of screening.
- Lack of evidence for harms associated with screening. Potential harms of a false positive result, or a true positive result when effective treatment is not available, include stigmatization and anxiety. Other than one study which reported a short-term increase in the number of unscheduled doctor visits in infants of screened women, we did not identify any studies that reported on outcomes for all women with positive results, not just those with a confirmatory diagnostic evaluation.
- Lack of evidence for an impact of screening and treatment of depression on longer term maternal and infant outcomes. This is ultimately needed to help in the weighting of harms versus benefits needed to help decide if, when, and whom to screen for postpartum depression. Although the consistent association between postpartum depression and a variety of adverse outcomes in infants and children is often cited as one of the primary rationales for screening, there is little or no direct evidence that screening and treatment leads to improved outcomes compared to no screening. Given that many of the social, relationship, and personality factors consistently associated with postpartum depression are also likely to be associated with suboptimal development outcomes in children, some evidence that treating depression in a single mother in a poor-quality relationship will lead to improved outcomes in children, even if the social factors do not change, would be helpful to strengthen the case for screening.
- Finally, one of the biggest barriers to synthesizing this literature is the diversity in research methods, definitions, and analytic tools used. Given the interdisciplinary nature of the condition, this diversity can be extremely helpful in bringing fresh insights to the problem. However, because of differences in preferred methods between fields, synthesis of results can be challenging. Even when the same technique is used, the results may be reported differently. For example, even though logistic regression is commonly used across a wide range of research as a method for multivariable analysis, different fields report the results differently. Medical and epidemiologic studies will report odds ratios

and confidence intervals, while some studies we reviewed in the psychological literature reported pseudo-R^2 values, or other summary statistics.

Research Gaps

General Gaps

Understanding the potential benefits and harms of screening for postpartum depression is an issue of considerable interest to patients, clinicians, and policymakers. Section 2952 of the 2010 Patient Protection Affordable Care Act provides for funding for research related to postpartum depression,[117] and there are two current funding opportunities from NIH specifically targeting mental health during pregnancy and the postpartum period.[118,119] This review has identified a number of research gaps that could be addressed utilizing these resources.

As noted above, one of the major limitations of the current evidence base is the wide disparity in methods and definitions used in studies relevant to screening for postpartum depression. This disparity limits the ability to synthesize the existing literature across disciplines; in particular, it significantly limits the ability to perform meta-analyses. It would be extremely valuable for researchers in the field to reach consensus on a core set of measures that would be reported consistently across all relevant studies. For studies of interventions, common outcomes measures are the highest priority. For observational studies, or other study designs where there is a need to adjust for potential confounding, common measures for both outcomes and confounders are needed. In practice, this means not only agreement on *which* variables to collect, but *how* to measure and report them. For example, parity is frequently reported as a mean and standard deviation, which is not only clinically meaningless (since non-integer values of number of deliveries have no interpretation), but does not reflect the underlying distribution.

For many of the recommendations below, use of formal simulation and decision models may prove useful. As illustrated above, even a simple model can be useful for illustrating trade-offs, and can highlight the relationship between uncertainty about the relative likelihood of adverse outcomes compared to favorable outcomes, the acceptable harm/benefit trade-off, and the extent to which further research will help clarify the optimal decision or recommendation. This approach can be done using both specific clinical outcomes, or it can explicitly incorporate costs; in the latter case, this value-of-information analysis can help inform research prioritization and research budgeting.[71,120] Further development of the model outlined in this report could incorporate variations in strategies, such as timing of screening relative to delivery, repeated screening at varying intervals during pregnancy and the postpartum period, use of strategies to target high risk groups for screening, and strategies to enhance followup and treatment of women with positive screening results.

KQ 1

- Although greater precision for sensitivity estimates would be useful, there will always be greater uncertainty about sensitivity than specificity in a screening setting, since the number of subjects with the underlying condition will always be much smaller than the number of subjects without the condition. Given this limitation, it would ultimately be more efficient to perform studies large enough to address the question directly rather than multiple additional smaller studies, particularly if the smaller studies focus on a single instrument. We would suggest

1. Achieving consensus on the appropriate trade-off between false positives and false negatives and using thresholds defined by these clinical criteria to determine optimal sensitivity and specificity for candidate screening instruments.
2. Determination of other criteria for evaluating screening instruments (ease of administration, time associated with administration, costs, patient and provider acceptability, etc.). These criteria could be collected as part of the study. Alternatively, patient and provider acceptability could be measured using methods such as discrete choice experiments to assess the relative importance of different attributes of the screening test;[121] these data could then be used to inform the choice of which instruments to evaluate further.
3. Define sample size for the study based on detecting clinically relevant differences in test performance and acceptability, with these differences being at least partially derived empirically in the first two steps.
4. Direct comparison of candidate instruments, either by having the same subject use each instrument (randomized as to order of administration), or randomizing different subjects to different instruments. The trade-off here is between the increased generalizability of having subjects take a single test versus overall sample size.

- The question of whether different instruments are better at identifying specific signs and symptoms is only important if there are effective interventions for those specific signs and symptoms. Clarity is needed on which signs and symptoms, and what potential interventions are available, in order to discuss potential research designs. One first step might be a systematic review focused on the individual signs and symptoms identified in the different screening instruments, with an emphasis on identifying effective interventions.

KQ 2

- Although we identified a number of consistent risk factors for postpartum depression, we did not identify any articles that used a multivariate predictive model to stratify patients by risk of developing the condition in order to screen more efficiently (similar to the Gail model used for identifying higher risk women for more aggressive screening protocols). The potential impact of such a model could be estimated by estimating the absolute risk of postpartum depression at different thresholds and using this information to estimate the number of false positives and false negatives resulting from screening only women identified as high risk. This could be compared to the estimated number of unwanted screening outcomes resulting from other strategies designed to minimize false positives, such as serial testing, using a simulation model. These data could, in turn, be used to estimate the size, costs, and value-of-information of a comparative trial.

KQs 3–6

- Additional RCTs comparing screening with no screening are needed. Ideally, some of these studies could address issues relevant to differences in timing, setting, or provider, perhaps through factorial designs.
- Explicit definitions of harms and benefits are needed.

    - Ideally, these studies should include a long-term followup component for both mothers and infants. Although this will substantially affect costs and timing of the studies, if the ultimate rationale for screening involves both maternal and child outcomes, then a more explicit demonstration of the benefits in terms of these longer term outcomes is needed.
- There was insufficient evidence to address the effect of timing, setting, or provider on test characteristics, while there was low strength evidence that timing might affect likelihood of receiving appropriate diagnostic and therapeutic services. It seems plausible that differences in outcome relevant to timing, setting, or provider are more related to aspects of the screening/referral/diagnosis process other than the test characteristics of the specific screening instrument used in the study.

Conclusions

Although currently available screening instruments are reasonably sensitive and specific in detecting postpartum depression, there is insufficient evidence to draw any conclusions about the net balance of benefits and harms of screening for postpartum depression, or about whether specific tools or strategies would result in a more favorable balance.

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR). Arlington, VA: American Psychiatric Association; 2000.

2. Jones I. DSMV: the perinatal onset specifier for mood disorders. Arlington, VA: American Psychiatric Association; 2010. Available at: http://www.dsm5.org/Documents/Mood%20Disorders%20Work%20Group/Ian%20Jones%20memo-post-partum.pdf. Accessed August 10, 2011.

3. Gaynes BN, Gavin N, Meltzer-Brody S, et al. Perinatal Depression: Prevalence, Screening Accuracy, and Screening Outcomes. Evidence Report/Technology Assessment No. 119 (Prepared by RTI–University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016). Rockville, MD: Agency for Healthcare Research and Quality; February 2005. AHRQ Publication No. 05-E006-2.

4. Gaynes BN, Gavin N, Meltzer-Brody S, et al. Perinatal depression: prevalence, screening accuracy, and screening outcomes. Evid Rep Technol Assess (Summ). 2005(119):1-8. PMID: 15760246.

5. Ng RC, Hirata CK, Yeung W, et al. Pharmacologic treatment for postpartum depression: a systematic review. Pharmacotherapy. 2010;30(9):928-41. PMID: 20795848.

6. Appleby L. Suicide during pregnancy and in the first postnatal year. BMJ. 1991;302(6769):137-40. PMID: 1995132.

7. Lindahl V, Pearson JL, Colpe L. Prevalence of suicidality during pregnancy and the postpartum. Archives of Women's Mental Health. 2005;8(2):77-87. PMID: 15883651.

8. Oates M. Perinatal psychiatric disorders: a leading cause of maternal morbidity and mortality. Br Med Bull. 2003;67:219-29. PMID: 14711766.

9. Spinelli MG. Maternal infanticide associated with mental illness: prevention and the promise of saved lives. Am J Psychiatry. 2004;161(9):1548-57. PMID: 15337641.

10. Chaudron LH, Szilagyi PG, Campbell AT, et al. Legal and ethical considerations: risks and benefits of postpartum depression screening at well-child visits. Pediatrics. 2007;119(1):123-8. PMID: 17200279.

11. Gjerdingen DK, Yawn BP. Postpartum depression screening: importance, methods, barriers, and recommendations for practice. J Am Board Fam Med. 2007;20(3):280-8. PMID: 17478661.

12. McLearn KT, Minkovitz CS, Strobino DM, et al. Maternal depressive symptoms at 2 to 4 months post partum and early parenting practices. Arch Pediatr Adolesc Med. 2006;160(3):279-84. PMID: 16520447.

13. Paris R, Bolton RE, Weinberg MK. Postpartum depression, suicidality, and mother-infant interactions. Arch Womens Ment Health. 2009;12(5):309-21. PMID: 19728036.

14. Petrou S, Cooper P, Murray L, et al. Economic costs of post-natal depression in a high-risk British cohort. Br J Psychiatry. 2002;181:505-12. PMID: 12456521.

15. Sills MR, Shetterly S, Xu S, et al. Association between parental depression and children's health care use. Pediatrics. 2007;119(4):e829-36. PMID: 17403826.

16. Anderson LN, Campbell MK, daSilva O, et al. Effect of maternal depression and anxiety on use of health services for infants. Can Fam Physician. 2008;54(12):1718-9 e5. PMID: 19074718.

17. Hewitt C, Gilbody S, Brealey S, et al. Methods to identify postnatal depression in primary care: an integrated evidence synthesis and value of information analysis. Health Technol Assess. 2009;13(36):1-145, 7-230. PMID: 19624978.

18. American College of Obstetricians and Gynecologists, Committee on Obstetric Practice. Committee opinion no. 453: Screening for depression during and after pregnancy. Obstet Gynecol. 2010;115(2 Pt 1):394-5. PMID: 20093921.

19. U. S. Preventive Services Task Force. Screening for depression in adults: U.S. preventive services task force recommendation statement. Ann Intern Med. 2009;151(11):784-92. PMID: 19949144.

20. Earls MF. Incorporating recognition and management of perinatal and postpartum depression into pediatric practice. Pediatrics. 2010;126(5):1032-9. PMID: 20974776.

21. Hirst KP, Moutier CY. Postpartum major depression. Am Fam Physician. 2010;82(8):926-33. PMID: 20949886.

22. American College of Nurse Midwives, Division of Women's Health Policy and Leadership. Position statement: depression in women; 2003. Available at: www.midwife.org/siteFiles/position/Depression_in_Women_05.pdf. Accessed August 10, 2011.

23. National Institute for Health and Clinical Excellence. Antenatal and Postnatal Mental Health: Clinical Management and Service Guidance. NICE Clinical Guideline 45. Developed by the National Collaborating Centre for Mental Health. London: National Institute for Health and Clinical Excellence; 2007. Available at: http://guidance.nice.org.uk/CG45/NICEGuidance/pdf/English. Accessed August 10, 2011.

24. Kozhimannil KB, Adams AS, Soumerai SB, et al. New Jersey's efforts to improve postpartum depression care did not change treatment patterns for women on Medicaid. Health Aff (Millwood). 2011;30(2):293-301. PMID: 21289351.

25. Barbui C, Cipriani A, Patel V, et al. Efficacy of antidepressants and benzodiazepines in minor depression: systematic review and meta-analysis. Br J Psychiatry. 2011;198(1):11-6. PMID: 21200071.

26. Eberhard-Gran M, Eskild A, Tambs K, et al. Depression in postpartum and non-postpartum women: prevalence and risk factors. Acta Psychiatr Scand. 2002;106(6):426-33. PMID: 12392485.

27. Josefsson A, Angelsiöö L, Berg G, et al. Obstetric, somatic, and demographic risk factors for postpartum depressive symptoms. Obstet Gynecol. 2002;99(2):223-8. PMID: 11814501.

28. McCoy SJ, Beal JM, Shipman SB, et al. Risk factors for postpartum depression: a retrospective investigation at 4-weeks postnatal and a review of the literature. J Am Osteopath Assoc. 2006;106(4):193-8. PMID: 16627773.

29. Robertson E, Grace S, Wallington T, et al. Antenatal risk factors for postpartum depression: a synthesis of recent literature. Gen Hosp Psychiatry. 2004;26(4):289-95. PMID: 15234824.

30. Vesga-Lopez O, Blanco C, Keyes K, et al. Psychiatric disorders in pregnant and postpartum women in the United States. Arch Gen Psychiatry. 2008;65(7):805-15. PMID: 18606953.

31. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed January 3, 2012.

32. Agency for Healthcare Research and Quality. Methods Guide for Medical Test Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=558. Accessed January 3, 2012.

33. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097. PMID: 19621072.

34. Anonymous. Evidence-based Practice Center Systematic Review Protocol. Project Title: Efficacy and Safety of Screening for Postpartum Depression. March 9, 2012. Available at: www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=997. Accessed June 21, 2012.

35. Austin MP, Priest SR, Sullivan EA. Antenatal psychosocial assessment for reducing perinatal mental health morbidity.

Cochrane Database Syst Rev. 2008(4):CD005124. PMID: 18843682.

36. Bennett HA, Einarson A, Taddio A, et al. Prevalence of depression during pregnancy: systematic review. Obstet Gynecol. 2004;103(4):698-709. PMID: 15051562.

37. Boyd RC, Le HN, Somberg R. Review of screening instruments for postpartum depression. Arch Womens Ment Health. 2005;8(3):141-53. PMID: 16133785.

38. Cuijpers P, Van Straten A, Warmerdam L, et al. Characteristics of effective psychological treatments of depression: a metaregression analysis. Psychother Res. 2008;18(2):225-36. PMID: 18815968.

39. DeBackere KJ, Hill PD, Kavanaugh KL. The parental experience of pregnancy after perinatal loss. J Obstet Gynecol Neonatal Nurs. 2008;37(5):525-37. PMID: 18811772.

40. Downe SM, Butler E, Hinder S. Screening tools for depressed mood after childbirth in UK-based South Asian women: a systematic review. J Adv Nurs. 2007;57(6):565-83. PMID: 17346316.

41. Gibson J, McKenzie-McHarg K, Shakespeare J, et al. A systematic review of studies validating the Edinburgh Postnatal Depression Scale in antepartum and postpartum women. Acta Psychiatr Scand. 2009;119(5):350-64. PMID: 19298573.

42. Gjerdingen D, Katon W, Rich DE. Stepped care treatment of postpartum depression: a primary care-based management model. Womens Health Issues. 2008;18(1):44-52. PMID: 18215764.

43. Goldbort J. Transcultural analysis of postpartum depression. MCN Am J Matern Child Nurs. 2006;31(2):121-6. PMID: 16523038.

44. Goodman JH. Postpartum depression beyond the early postpartum period. J Obstet Gynecol Neonatal Nurs. 2004;33(4):410-20. PMID: 15346666.

45. Halbreich U, Karkun S. Cross-cultural and social diversity of prevalence of postpartum depression and depressive symptoms. J Affect Disord. 2006;91(2-3):97-111. PMID: 16466664.

46. Hewitt CE, Gilbody SM. Is it clinically and cost effective to screen for postnatal depression: a systematic review of controlled clinical trials and economic evidence. BJOG. 2009;116(8):1019-27. PMID: 19438499.

47. Klainin P, Arthur DG. Postpartum depression in Asian cultures: a literature review. Int J Nurs Stud. 2009;46(10):1355-73. PMID: 19327773.

48. Kroenke K, Spitzer RL, Williams JB, et al. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen Hosp Psychiatry. 2010;32(4):345-59. PMID: 20633738.

49. Leddy MA, Lawrence H, Schulkin J. Obstetrician-gynecologists and women's mental health: findings of the Collaborative Ambulatory Research Network 2005-2009. Obstet Gynecol Surv. 2011;66(5):316-23. PMID: 21794195.

50. Mann R, Gilbody S. Validity of two case finding questions to detect postnatal depression: a review of diagnostic test accuracy. J Affect Disord. 2011;133(3):388-97. PMID: 21146230.

51. Mian AI. Depression in pregnancy and the postpartum period: balancing adverse effects of untreated illness with treatment risks. J Psychiatr Pract. 2005;11(6):389-96. PMID: 16304507.

52. Minkovitz CS, Strobino D, Scharfstein D, et al. Maternal depressive symptoms and children's receipt of health care in the first 3 years of life. Pediatrics. 2005;115(2):306-14. PMID: 15687437.

53. O'Mahony J, Donnelly T. Immigrant and refugee women's post-partum depression help-seeking experiences and access to care: a review and analysis of the literature. J Psychiatr Ment Health Nurs. 2010;17(10):917-28. PMID: 21078007.

54. Ross LE, Campbell VL, Dennis CL, et al. Demographic characteristics of participants in studies of risk factors, prevention, and treatment of postpartum depression. Can J Psychiatry. 2006;51(11):704-10. PMID: 17121169.

55. Shaw E, Levitt C, Wong S, et al. Systematic review of the literature on postpartum care: effectiveness of postpartum support to improve maternal parenting, mental health,

quality of life, and physical health. Birth. 2006;33(3):210-20. PMID: 16948721.

56. Wylie L, Hollins Martin CJ, Marland G, et al. The enigma of post-natal depression: an update. J Psychiatr Ment Health Nurs. 2011;18(1):48-58. PMID: 21214684.

57. Yozwiak JA. Postpartum depression and adolescent mothers: a review of assessment and treatment approaches. J Pediatr Adolesc Gynecol. 2010;23(3):172-8. PMID: 20496498.

58. Zubaran C, Schumacher M, Roxo MR, et al. Screening tools for postpartum depression: validity and cultural dimensions. Afr J Psychiatry (Johannesbg). 2010;13(5):357-65. PMID: 21390406.

59. Brealey SD, Hewitt C, Green JM, et al. Screening for postnatal depression—Is it acceptable to women and healthcare professionals? A systematic review and meta-synthesis. Journal of Reproductive and Infant Psychology. 2010;28(4):328-44. PMID: 2010-24613-003.

60. Hewitt CE, Gilbody SM, Mann R, et al. Instruments to identify post-natal depression: Which methods have been the most extensively validated, in what setting and in which language? International Journal of Psychiatry in Clinical Practice. 2010;14(1):72-6. PMID: 2010-12805-012.

61. Matthey S. Detection and treatment of postnatal depression (perinatal depression or anxiety). Current Opinion in Psychiatry. 2004;17(1):21-9. PMID: 2004-18915-005.

62. Brealey SD, Hewitt C, Green JM, et al. Screening for postnatal depression - is it acceptable to women and healthcare professionals? A systematic review and meta-synthesis. Journal of Reproductive and Infant Psychology. 2010;28(4):328-44.

63. Farneti C, Farneti M. Review of postpartum depression: A pathology also for pediatricians? Acta Medica Mediterranea. 2009;25(3):153-61.

64. The World Bank. Country and lending groups. Available at: http://data.worldbank.org/about/country-classifications/country-and-lending-groups. Accessed January 3, 2012.

65. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36. PMID: 22007046.

66. Harbord RM, Whiting P, Sterne JA, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol. 2008;61(11):1095-103. PMID: 19208372.

67. Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. Methods Inf Med. 2010;49(1):54-62, -4. PMID: 19936437.

68. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60. PMID: 12958120.

69. Myers ER, Misurski DA, Swamy GK. Influence of timing of seasonal influenza vaccination on effectiveness and cost-effectiveness in pregnancy. Am J Obstet Gynecol. 2011;204(6 Suppl 1):S128-40. PMID: 21640230.

70. Ungar WJ, ed. Economic Evaluation in Child Health. New York: Oxford University Press; 2010.

71. Myers E, Sanders GD, Ravi D, et al. Evaluating the Potential Use of Modeling and Value-of-Information Analysis for Future Research Prioritization Within the Evidence-based Practice Center Program. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-2007-10066-I.) AHRQ Publication No. 11-EHC030-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed January 3, 2012.

72. Owens DK, Lohr KN, Atkins D, et al. AHRQ Series Paper 5: Grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the Effective Health-Care Program. J Clin Epidemiol. 2010;63(5):513-23. PMID: 19595577.

73. Austin MP, Colton J, Priest S, et al. The Antenatal Risk Questionnaire (ANRQ): Acceptability and use for psychosocial risk assessment in the maternity setting. Women Birth. 2011. PMID: 21764399.

74. Austin MP, Hadzi-Pavlovic D, Priest SR, et al. Depressive and anxiety disorders in the postpartum period: how prevalent are they and can we improve their detection? Arch Womens Ment Health. 2010;13(5):395-401. PMID: 20232218.

75. Beck CT, Gable RK. Screening performance of the postpartum depression screening scale—Spanish version. J Transcult Nurs. 2005;16(4):331-8. PMID: 16160195.

76. Chaudron LH, Szilagyi PG, Tang W, et al. Accuracy of depression screening tools for identifying postpartum depression among urban mothers. Pediatrics. 2010;125(3):e609-17. PMID: 20156899.

77. Clarke PJ. Validation of two postpartum depression screening scales with a sample of First Nations and Metis women. Can J Nurs Res. 2008;40(1):113-25. PMID: 18459275.

78. Csatordai S, Kozinszky Z, Devosa I, et al. Validation of the Leverton Questionnaire as a screening tool for postnatal depression in Hungary. Gen Hosp Psychiatry. 2009;31(1):56-66. PMID: 19134511.

79. Edmondson OJ, Psychogiou L, Vlachos H, et al. Depression in fathers in the postnatal period: assessment of the Edinburgh Postnatal Depression Scale as a screening measure. J Affect Disord. 2010;125(1-3):365-8. PMID: 20163873.

80. Felice E, Saliba J, Grech V, et al. Validation of the Maltese version of the Edinburgh Postnatal Depression Scale. Arch Womens Ment Health. 2006;9(2):75-80. PMID: 16172837.

81. Gjerdingen D, Crow S, McGovern P, et al. 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. PMID: 19139451.

82. Hamdan A, Tamim H. Psychosocial risk and protective factors for postpartum depression in the United Arab Emirates. Arch Womens Ment Health. 2011;14(2):125-33. PMID: 21063891.

83. Howard LM, Flach C, Mehay A, et al. The prevalence of suicidal ideation identified by the Edinburgh Postnatal Depression Scale in postpartum women in primary care: findings from the RESPOND trial. BMC Pregnancy Childbirth. 2011;11:57. PMID: 21812968.

84. Jardri R, Pelta J, Maron M, et al. Predictive validation study of the Edinburgh Postnatal Depression Scale in the first week after delivery and risk analysis for postnatal depression. J Affect Disord. 2006;93(1-3):169-76. PMID: 16644021.

85. 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):213-9. PMID: 20542520.

86. Navarro P, Ascaso C, Garcia-Esteve L, et al. Postnatal psychiatric morbidity: a validation study of the GHQ-12 and the EPDS as screening tools. Gen Hosp Psychiatry. 2007;29(1):1-7. PMID: 17189737.

87. Pereira AT, Bos S, Marques M, et al. The Portuguese version of the postpartum depression screening scale. J Psychosom Obstet Gynaecol. 2010;31(2):90-100. PMID: 20443658.

88. Akincigil A, Munch S, Niemczyk KC. Predictors of maternal depression in the first year postpartum: marital status and mediating role of relationship quality. Soc Work Health Care. 2010;49(3):227-44. PMID: 20229395.

89. Andersson L, Sundstrom-Poromaa I, Wulff M, et al. Depression and anxiety during pregnancy and six months postpartum: a follow-up study. Acta Obstet Gynecol Scand. 2006;85(8):937-44. PMID: 16862471.

90. Barnes J, Senior R, MacPherson K. The utility of volunteer home-visiting support to prevent maternal depression in the first year of life. Child Care Health Dev. 2009;35(6):807-16. PMID: 19719770.

91. Beck CT, Froman RD, Bernal H. Acculturation level and postpartum depression in Hispanic mothers. MCN Am J Matern Child Nurs. 2005;30(5):299-304. PMID: 16132006.

92. Bloch M, Rotenberg N, Koren D, et al. Risk factors associated with the development of postpartum mood disorders. J Affect Disord. 2005;88(1):9-18. PMID: 15979150.

93. Boyce P, Hickey A. Psychosocial risk factors to major depression after childbirth. Soc Psychiatry Psychiatr Epidemiol. 2005;40(8):605-12. PMID: 16096700.

94. Chee CY, Chong YS, Ng TP, et al. The association between maternal depression and frequent non-routine visits to the infant's doctor—a cohort study. J Affect Disord. 2008;107(1-3):247-53. PMID: 17869346.

95. Chee CY, Lee DT, Chong YS, et al. Confinement and other psychosocial factors in perinatal depression: a transcultural study in Singapore. J Affect Disord. 2005;89(1-3):157-66. PMID: 16257451.

96. Garcia-Esteve L, Navarro P, Ascaso C, et al. Family caregiver role and premenstrual syndrome as associated factors for postnatal depression. Arch Womens Ment Health. 2008;11(3):193-200. PMID: 18506575.

97. Mauri M, Oppo A, Montagnani MS, et al. Beyond "postpartum depressions": specific anxiety diagnoses during pregnancy predict different outcomes: results from PND-ReScU. J Affect Disord. 2010;127(1-3):177-84. PMID: 20554326.

98. Turner K, Piazzini A, Franza A, et al. Epilepsy and postpartum depression. Epilepsia. 2009;50 Suppl 1:24-7. PMID: 19125843.

99. Verkerk GJ, Denollet J, Van Heck GL, et al. Personality factors as determinants of depression in postpartum women: a prospective 1-year follow-up study. Psychosom Med. 2005;67(4):632-7. PMID: 16046379.

100. Kersting A, Kroker K, Steinhard J, et al. Complicated grief after traumatic loss: A 14-month follow up study. Eur Arch Psychiatry Clin Neurosci. 2007;257(8):437-43. PMID: 2008-00696-002.

101. Crotty F, Sheehan J. Prevalence and detection of postnatal depression in an Irish community sample. Ir J Psychol Med. 2004;21(4):117-21. PMID: 2004-22274-003.

102. Zlotnick C, Miller IW, Pearlstein T, et al. A preventive intervention for pregnant women on public assistance at risk for postpartum depression. Am J Psychiatry. 2006;163(8):1443-5. PMID: 16877662.

103. Leung SS, Leung C, Lam TH, et al. Outcome of a postnatal depression screening programme using the Edinburgh Postnatal Depression Scale: a randomized controlled trial. J Public Health (Oxf). 2011;33(2):292-301. PMID: 20884642.

104. Morrell CJ, Slade P, Warner R, et al. Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: Pragmatic cluster randomised trial in primary care. BMJ. 2009;338(7689):276-9.

105. Glavin K, Smith L, Sorum R, et al. Redesigned community postpartum care to prevent and treat postpartum depression in women—a one-year follow-up study. J Clin Nurs. 2010;19(21-22):3051-62. PMID: 20726926.

106. Cooper PJ, Murray L, Hooper R, et al. The development and validation of a predictive index for postpartum depression. Psychol Med. 1996;26(3):627-34. PMID: 8733220.

107. Goodman JH, Tyer-Viola L. Detection, treatment, and referral of perinatal depression and anxiety by obstetrical providers. J Womens Health (Larchmt). 2010;19(3):477-90. PMID: 20156110.

108. Ballestrem CL, Strauss M, Kachele H. Contribution to the epidemiology of postnatal depression in Germany—implications for the utilization of treatment. Arch Womens Ment Health. 2005;8(1):29-35. PMID: 15868391.

109. Burton A, Patel S, Kaminsky L, et al. Depression in pregnancy: time of screening and access to psychiatric care. J Matern Fetal Neonatal Med. 2011;24(11):1321-4. PMID: 21261444.

110. Flynn HA, O'Mahen HA, Massey L, et al. The impact of a brief obstetrics clinic-based intervention on treatment use for perinatal depression. J Womens Health (Larchmt). 2006;15(10):1195-204. PMID: 17199460.

111. Yonkers KA, Smith MV, Lin H, et al. Depression screening of perinatal women: an evaluation of the healthy start depression initiative. Psychiatr Serv. 2009;60(3):322-8. PMID: 19252044.

112. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012;132(2):365-77. PMID: 22037780.

113. Atkins D, Chang SM, Gartlehner G, et al. Assessing applicability when comparing medical interventions: AHRQ and the

Effective Health Care Program. J Clin Epidemiol. 2011;64(11):1198-207. PMID: 21463926.

114. Genders TS, Meijboom WB, Meijs MF, et al. CT coronary angiography in patients suspected of having coronary artery disease: decision making from various perspectives in the face of uncertainty. Radiology. 2009;253(3):734-44. PMID: 19864509.

115. Havrilesky LJ, Sanders GD, Kulasingam S, et al. Development of an ovarian cancer screening decision model that incorporates disease heterogeneity: implications for potential mortality reduction. Cancer. 2011;117(3):545-53. PMID: 21254049.

116. 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):187-93. PMID: 21383219.

117. Office of the Legislative Counsel for the use of the U.S. House of Representatives. Patient Protection and Affordable Care Act. Text available at: www.healthcare.gov/law/full. Accessed July 25, 2012.

118. National Institutes of Health. Women's Mental Health During Pregnancy and the Postpartum Period (R01 Research Project Grant). Available at: http://grants.nih.gov/grants/guide/pa-files/PA-12-216.html. Accessed July 25, 2012.

119. National Insitutes of Health. Women's Mental Health During Pregnancy and the Postpartum Period (R21 Exploratory/Developmental Research Grant). Available at: http://grants.nih.gov/grants/guide/pa-files/PA-12-215.html. Accessed July 25, 2012.

120. Myers E, McBroom AJ, Shen L, et al. Value-of-Information Analysis for Patient-Centered Outcomes Research Prioritization. Prepared by the Duke Evidence-based Practice Center for the Patient-Centered Outcomes Research Institute. March 9, 2012. Available at: www.pcori.org/assets/Value-of-Information-Analysis-for-Patient-Centered-Outcomes-Research-Prioritization.pdf. Accessed July 3, 2012.

121. Wordsworth S, Ryan M, Skatun D, et al. Women's preferences for cervical cancer screening: a study using a discrete choice experiment. Int J Technol Assess Health Care. 2006;22(3):344-50. PMID: 16984063.

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
BDI	Beck Depression Inventory
BDI-IA	Beck Depression Inventory-IA
BDI-II	Beck Depression Inventory-II
BPDS	Bromley Postnatal Depression Scale
BSID	Bayley Scales of Infant Development
CDSR	Cochrane Database of Systematic Reviews
CER	Comparative Effectiveness Review
CES-D	Center for Epidemiologic Studies Depression Scale
CI	confidence interval
DIS	Diagnostic Interview Schedule
DSM-IV-TR	*Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision*
DSM-V	*Diagnostic and Statistical Manual of Mental Disorders, 5th Edition*
EPC	Evidence-based Practice Center
EPDS	Edinburgh Postnatal Depression Scale
GHQ-D	General Health Questionnaire
GHQ-12	12-Item General Health Questionnaire
HADS	Hospital Anxiety and Depression Scale
HRSD	Hamilton Rating Scale for Depression
HRSD-17	17-Item Hamilton Rating Scale for Depression
HRSD-21	21-Item Hamilton Rating Scale for Depression
ICD	International Classification of Diseases
ICTRP	International Clinical Trials Registry Platform
ISMI	Internalized Stigma of Mental Illness
KQ	Key Question
LQ	Leverton Questionnaire
LR-	negative likelihood ratio
LR+	positive likelihood ratio
MADRS	Montgomery Asberg Depression Rating Scale
NPV	negative predictive value
OR	odds ratio
PDPI-R	Postpartum Depression Predictors Inventory-Revised
PDSS	Postpartum Depression Screening Scale
PHQ-2	2-Item Patient Health Questionnaire
PHQ-9	9-Item Patient Health Questionnaire
PICOTS	Populations, Interventions, Comparators, Outcomes, Timings, and Settings of interest
PDSS	Postpartum Depression Screening Scale
PP-E	early postpartum period;
PP-L	late postpartum period
PPV	positive predictive value
PRIME-MD CEG	Primary Care Evaluation of Mental Disorders Clinical Evaluation Guide

PRIME-MD PHQ	Primary Care Evaluation of Mental Disorders Patient Health Questionnaire
PRIME-MD PQ	Primary Care Evaluation of Mental Disorders Patient Questionnaire
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSI	Parenting Stress Index
QALY	quality-adjusted life year
QUADAS-2	QUality Assessment of Diagnostic Accuracy Studies-2
RCT	randomized controlled trial
RDC	Research Diagnostic Criteria
ROC	receiver operating characteristic
RR	relative risk
SADS	Schedule for Affective Disorders and Schizophrenia
SCAN	Schedules for Clinical Assessment in Neuropsychiatry
SCID	Structured Clinical Interview for Depression
SF-36	Medical Outcomes Study 36-Item Short-Form Health Survey
SPI	Goldberg's Standardized Psychiatric Interview
TEP	Technical Expert Panel
TOO	Task Order Officer
USPSTF	U.S. Preventive Services Task Force
WHO	World Health Organization
Zung SDS	Zung Self-Rating Depression Scale