Comparative Effectiveness Review
Number 106

Effective Health Care Program

Efficacy and Safety of Screening for Postpartum Depression

Executive Summary

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 of depression in other populations. Postpartum depression is defined in the "Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision" (DSM-IV-TR)[1] as a major depressive disorder according to standard diagnostic criteria—namely, five 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:

- 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, weight gain (e.g., change of more than 5 percentof 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)

Effective Health Care Program

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**.

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Effective
Health Care

- 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

A new set of diagnostic criteria for psychiatric illness, the "Diagnostic and Statistical Manual of Mental Disorders, 5th Edition" (DSM-5), is currently scheduled for release in May 2013.

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 (two to four 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 formal synthesis of the evidence, performed for the Agency for Healthcare Research and Quality (AHRQ) in 2005,[2,3] estimated that the point prevalence (the proportion of the population with the condition at a given point in time) of major depression alone during the first postpartum year is 1.0–5.9 percent, with point prevalence for major and minor depression combined of 6.5–12.9 percent. The AHRQ evidence review found a best estimate for period prevalence (the proportion of the population with the condition at any point during a defined time period) of 21.9 percent (95% confidence interval [CI], 15.1 to 30.0%).[3] Incidence (the rate of new cases among a population without the condition within a given time period) 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, with a cumulative 12-month incidence of 30.6 percent (95% CI, 18.3 to 45.4%). Although depression in the perinatal period has attracted special interest, the available data suggest that incidence and prevalence of major depression in the postpartum period are comparable to rates observed in women of reproductive age who are not pregnant or postpartum. However, the prevalence of depressive symptoms not meeting diagnostic criteria for depression may be higher, particularly in the first 3 months after birth.[3,5] Depression in adults has a significant impact on quality of life, productivity, and social functioning,[5,6] and there is no evidence that these effects are any different for women during the postpartum period. Mortality is also a risk for mothers through suicide and for infants through neglect, abuse, or homicide. As noted in a 2009 report by the Institute of Medicine,[5] 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, some of which may be inappropriate, 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. Screening can potentially improve outcomes by identifying undiagnosed depression that would otherwise either go untreated or be treated at a more severe stage. There is universal recognition of the harms associated with 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 and 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 persistent uncertainty about how well currently available tests and strategies perform in identifying women who may have, or are at risk for, postpartum depression. It is also uncertain (1) how factors such as timing relative to delivery, setting, and provider might affect the performance of these strategies and (2) which factors 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 among different screening strategies.

Scope and Key Questions

This comparative effectiveness review (CER) was funded by AHRQ and designed to evaluate the comparative diagnostic accuracy, benefits, and harms of available screening instruments for postpartum depression. As specified in the Key Questions, we further considered whether the diagnostic accuracy, benefits, and harms of the screening instruments evaluated differed among specific patient subgroups of interest, defined by any of the following factors: age, race/ethnicity, parity, history of mood disorders, history of intimate partner violence, perinatal outcomes, or cultural factors. We also considered whether the performance characteristics of screening instruments were affected by the timing of screening, the setting in which screening was conducted, or the type of provider. This review does not consider questions regarding the safety and/or effectiveness of downstream options for postpartum depression treatment. Treatment options are being addressed in another AHRQ CER (currently in progress) that will be published as a separate report.

By summarizing the available evidence on the accuracy and effectiveness of screening for postpartum depression, we hope to provide a resource to organizations developing recommendations to enhance patient-centered outcomes for women, their partners, and children, ideally with efficient use of clinical resources. We also identify key areas of uncertainty that limit stakeholders' ability to adequately judge the balance of benefits and harms associated with screening at both the individual and system level, and suggest areas where additional research to specifically address the limitations of the currently available evidence would help resolve this uncertainty.

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 with 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 with 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 CER follow those suggested in the AHRQ "Methods Guide for Effectiveness and Comparative Effectiveness Reviews" (Methods Guide)[7] and "Methods Guide for Medical Test Reviews" (Medical Test Guide).[8]

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 on the AHRQ Effective Health Care Web site.[9]

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 July 24, 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 gray 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. Gray 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. During peer and public review of the draft report, we updated all database searches and included any eligible studies identified either through that search or through suggestions from peer and public reviewers.

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 full report. For all KQs, the search focused on studies that were conducted in economically developed countries, were 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. We focused on economically developed countries, which have greater cultural and health care system similarities to the United States, to improve the applicability of the review findings to U.S. populations. 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 reporting depression outcomes were required to include confirmation of depression with a reference standard. 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] vs. 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[7] and Medical Test Guide.[8] 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 for KQs 1, 2, and 3, we used QUADAS-2 (QUality Assessment of Diagnostic Accuracy Studies-2[10]) 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 (in terms of both study population and outcomes), and completeness of the reporting of results. We considered random-effects meta-analyses for comparisons where at least three conceptually homogeneous studies reported the same patient-centered intermediate, final, or adverse effect outcome. Test performance was summarized using sensitivity and specificity. Where 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 (three or more), we planned 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.[11] 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 of treated depression (true positives), false negatives, and false positives as model output, and multiplied these probabilities by 4 million (the approximate annual number of deliveries in the United States) to estimate the number of women likely to experience these outcomes under different screening approaches. 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 that 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.[12] 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[7,13] and Medical Test Guide.[8] 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[7,13] and the Medical Test Guide.[8] In brief, this method uses the PICOTS (populations, interventions, comparators, outcomes, timing, and settings) 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., whether a screening test was used in the general population vs. in a subgroup preselected based on known or suspected risk factors), family income, race, ethnicity, parity, and partner support. Within this report we consider studies conducted in the United Kingdom (UK) separately from those conducted in the rest of Europe, primarily because the use of screening instruments administered in English enhances the applicability of UK studies to a U.S. nonimmigrant setting. 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 with 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 KQ. For 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 5,059 citations, 1,528 of which were duplicate citations. Manual searching identified 154 additional citations, for a total of 3,685 citations to be screened. After applying inclusion/exclusion criteria at the title-and-abstract level, 1,293 full-text articles were retrieved and screened. Of these, 1,248 were excluded at the full-text screening stage, leaving 45 articles for data abstraction. These 45 articles described 40 unique studies. The relationship of studies to the review questions is as follows: 18 studies relevant to KQ 1, 15 studies relevant to

KQ 2, 2 studies relevant to KQ 3, 5 studies relevant to KQ 4, 1 study relevant to KQ 5, and 6 studies relevant to KQ 6. (Some studies were relevant to more than one KQ.)

KQ 1. Performance Characteristics of Screening Instruments

We identified 18 studies (1 of which focused on fathers) 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; six in Europe; four in the UK; and one each in Australia, New Zealand, Asia, and Canada. Ten 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 in which two or more screening tests were directly compared. Only one study was relevant to KQ 1b.

Eleven studies provided data on the Edinburgh Postnatal Depression Scale (EPDS), four on the Postpartum Depression Screening Scale (PDSS), four on various versions of the Beck Depression Inventory (BDI), two on a "two-question" screen, and one each on the Patient Health Questionnaire (PHQ-9), the Antenatal Risk Questionnaire, the 17- and 21-Item Hamilton Rating Scale for Depression (HRSD-17 and HRSD-21), and the Leverton Questionnaire.

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. A "yes" response to either of the questions in the two-question screen had sensitivity of 100 percent in two studies, with specificities of 44.5 and 65.7 percent. Because of the heterogeneity among studies in terms of setting, population, and choice of screening threshold, we were unable to perform quantitative synthesis, and CIs between tests broadly overlapped.

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

Screening Test	Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Test Performance (95% CI)
			Risk of Bias	Consistency	Directness	Precision	
Antenatal Risk Questionnaire	Sensitivity	1 (276)	High	NA	Direct	Imprecise	Low SOE 78.1% (65.0–88.7%)
	Specificity	1 (276)	High	NA	Direct	Imprecise	Low SOE 47.1% (40.3–59.9%)
BDI	Sensitivity	2 (1,151)	Medium	Consistent	Direct	Imprecise	Low SOE 80–90% (approximate range of point estimates at most commonly used thresholds)
	Specificity	2 (1,151)	Medium	Consistent	Direct	Precise	Low SOE 80–90% (approximate range of point estimates at most commonly used thresholds)

Table A. Strength-of-evidence domains for test characteristics of screening tests for postpartum depression (continued)

Screening Test	Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Test Performance (95% CI)
			Risk of Bias	Consistency	Directness	Precision	
BDI-II	Sensitivity	2 (650)	Medium	Consistent	Direct	Imprecise	Low SOE 75–90% (approximate range of point estimates at most commonly used thresholds)
	Specificity	2 (650)	Medium	Consistent	Direct	Precise	Low SOE 80–90% (approximate range of point estimates at most commonly used thresholds)
EPDS	Sensitivity	11 (3,456)	Medium	Consistent	Direct	Imprecise	Moderate SOE 80–90% (approximate range of point estimates at most commonly used thresholds)
	Specificity	11 (3,456)	Medium	Consistent	Direct	Precise	Moderate SOE 80–90% (approximate range of point estimates at most commonly used thresholds)
HRSD-17	Sensitivity	1 (534)	High	NA	Direct	Imprecise	Low SOE 80–85% (range of point estimates across thresholds)
	Specificity	1 (534)	High	NA	Direct	Imprecise	Low SOE 80–85% (range of point estimates across thresholds)
HRSD-21	Sensitivity	1 (534)	High	NA	Direct	Imprecise	Low SOE 80–85% (range of point estimates across thresholds)
	Specificity	1 (534)	High	NA	Direct	Imprecise	Low SOE 75–80% (range of point estimates across thresholds)

Table A. Strength-of-evidence domains for test characteristics of screening tests for postpartum depression (continued)

Screening Test	Outcome	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Test Performance (95% CI)
			Risk of Bias	Consistency	Directness	Precision	
Leverton Questionnaire	Sensitivity	1 (617)	Low	NA	Direct	Imprecise	Low SOE 95.2% (90.4–98.1%)
	Specificity	1 (617)	Low	NA	Direct	Imprecise	Low SOE 91.3% (88.4–93.7%)
PDSS	Sensitivity	4 (903)	Medium	Consistent	Direct	Imprecise	Moderate SOE 80–90% (approximate range of point estimates at most commonly used thresholds)
	Specificity	4 (903)	Medium	Consistent	Direct	Precise	Moderate SOE 80–90% (approximate range of point estimates at most commonly used thresholds)
PHQ-9	Sensitivity	1 (506)	Low	NA	Direct	Imprecise	Low SOE 75–89% (range of point estimates at varying thresholds; 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 thresholds)
Two-Question Screen	Sensitivity	2 (600)	Low	Consistent	Direct	Imprecise	Moderate SOE 100% (sensitivity 100% in both studies)
	Specificity	2 (600)	Low	Consistent	Direct	Imprecise	Moderate SOE 44.3–65.7%

BDI = Beck Depression Inventory; BDI-II = Beck Depression Inventory-II; 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; SOE = strength of evidence

KQ 2. Effect of Individual Factors on Screening Performance

We identified 16 articles describing 15 unique studies that met the inclusion criteria for KQ 2. Three were from the United States; seven were from Europe; two were from Asia; and there was one study each from the UK, Australia, and Israel. Two studies were rated low risk of bias, 10 high risk of bias, and 3 unclear risk of bias. We did not identify any studies relevant to KQ 2b. Only one study judged to be at high risk of bias provided a specific estimate of the effect of a risk factor on test characteristics. 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	3 (5,578)	Medium	Inconsistent	Direct	Imprecise	Insufficient
	Education	2 (4,757)	Medium	Inconsistent	Direct	Imprecise	Insufficient
	Income	1 (4,245)	Medium	NA	Direct	Imprecise	Insufficient
	Employment status (unemployed vs. employed)	1 (363)	High	NA	Direct	Imprecise	Low SOE for increased risk of postpartum depression in unemployed mothers OR, 2.8 (1.1–4.9)
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 (4,998)	Medium	Inconsistent	Direct	Imprecise	Insufficient
	Alcohol use	1 (4,348)	Medium	NA	Direct	Imprecise	Insufficient
General Medical History	Poor health status/chronic illness	2 (4,993)	Medium	Consistent	Direct	Imprecise	Low SOE for increased risk of postpartum depression
	Obesity	1 (598)	Medium	NA	Direct	Imprecise	Insufficient
Psychiatric History	History of perinatal depression	2 (1,082)	High	Consistent	Direct	Imprecise	Low SOE for increased risk of postpartum depression
	History of depression	5 (2,057)	Medium	Consistent	Direct	Precise	Moderate SOE for increased risk of postpartum depression

Table B. Strength-of-evidence domains for associations with patient characteristics and risk of postpartum depression (continued)							
Risk Factor		Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
			Risk of Bias	Consistency	Directness	Precision	
Psychiatric History (continued)	History of premenstrual dysphoric disorder	1 (210)	Medium	NA	Direct	Imprecise	Low SOE for increased risk of postpartum depression
	Any psychiatric diagnosis	2 (1,075)	Medium	Consistent	Direct	Imprecise	Low SOE for increased risk of postpartum depression
	Anxiety	2 (1,305)	Medium	Consistent	Direct	Imprecise	Low SOE for increased risk of postpartum depression
	Personality (vulnerable/ neuroticism)	2 (685)	Medium	Consistent	Direct	Imprecise	Low SOE for increased risk of postpartum depression
Relationship/ Social Support	Marital status (single/no relationship)	3 (5,803)	Medium	Consistent	Direct	Imprecise	Low SOE for increased risk of postpartum depression
	Poor relationship quality	5 (6,101)	Medium	Consistent	Direct	Imprecise	Moderate SOE for increased risk of postpartum depression
	Poor social support	4 (1,830)	Medium	Consistent	Direct	Imprecise	Moderate SOE for increased risk of postpartum depression

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 postpartum depression and maternal age, education, income, or type of employment. One study did, however, find a significant association between maternal unemployment and postpartum depression (odds ratio [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 second or third 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 who did not terminate the pregnancy and whose children had severe abnormalities.

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 twofold.

Past history of depression or anxiety, including both postpartum and before pregnancy, were consistently associated with an increased risk of postpartum depression, with ORs 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 the 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.

Although the presence of any of these risk factors would presumably improve the positive predictive value of screening, only one study specifically reported on test characteristics stratified by individual patient characteristics; sensitivity of both the BDI and EPDS was lower in multigravid women compared with primigravid, but CIs were wide and overlapping.

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. Neither a U.S.-based study of two self-administered tests (BDI, EPDS) and two clinician-administered tests (HSRD-17, HSRD-21) nor an Irish-based study of the EPDS identified a significant effect of timing on test characteristics (Table C).

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

Five studies met our inclusion criteria and evaluated the comparative benefits of screening for postpartum depression. Four were RCTs, and one was a quasi-experimental study. Of the four RCTs, one was judged poor quality, two fair, and one good quality. 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 a single site to 4,084 enrolled from 101 practices. Two studies were conducted in the United States, and the others were conducted in the UK, Norway, and Hong Kong. Only the study conducted in Hong Kong provided any evidence regarding harms.

Table D summarizes the strength of evidence and findings. Three studies directly compared organized screening with no screening or "usual care." One fair-quality RCT found improvement in EPDS scores at 6 months in women randomized to screening at 2 months postdelivery compared with 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 (Medical Outcomes Study 12-Item Short-Form Health Survey) at 12 and 18 months in women randomized to screening, but no differences in other outcomes. A fair-quality U.S.-based study of primary care practices where screening, diagnosis, and treatment were carried out in the same practice found significant decreases in depression scores among the screened group, with rates of diagnosis substantially higher than those reported in other studies. None of the studies (the quasi-experimental study, the two fair-quality RCTs, and the one poor-quality RCT) that included the Parental Stress Inventory (PSI) or PSI-Short Form (PSI-SF) as an outcome showed a significant improvement in PSI scores with screening and treatment, despite showing improvement in depressive symptoms.

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

Six 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, one was a quasi-experimental design, and one was an RCT in

Table C. Strength-of-evidence domains for the effect of varying timing on screening for postpartum depression

Timing	Number of Studies (Subjects)	Domains Pertaining to SOE				SOE and Magnitude of Effect (95% CI)
		Risk of Bias	Consistency	Directness	Precision	
Delivery to 8 weeks vs. 8 weeks to 6 months	1 (534)	High	NA	Direct	Imprecise	Insufficient
Delivery vs. 6 weeks	1 (113)	High	NA	Direct	Imprecise	Insufficient

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

Table D. Strength-of-evidence domains for benefits and harms of screening for postpartum depression							
			Domains Pertaining to SOE				
Benefits/ Harms	Outcome	Number of Studies (Subjects)	Risk of Bias	Consistency	Directness	Precision	SOE and Magnitude of Effect (95% CI)
Benefits	Depressive symptoms	5 (8,071)	Medium	Consistent	Direct	Imprecise	Low to moderate SOE for reduced number of symptoms with screening and intervention
	Mental health score (SF-12)	1 (2,579)	Low	NA	Direct	Imprecise	Low SOE for improved scores with screening and intervention
	Parental stress	4 (5,567)	Medium	Consistent	Direct	Imprecise	Low SOE for no improvement in parental stress with screening and intervention
Harms	Unscheduled doctor visits for infant	1 (462)	Medium	NA	Direct	Imprecise	Low SOE for increased number of visits for infants of screened women

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

which randomization was performed at the primary care practice level. One cohort study was rated as fair quality and one was poor quality. The cross-sectional study was rated as good quality, the pre-post intervention study and quasi-experimental study were rated as poor quality, and the RCT was rated as fair quality. All six studies were conducted in the United States. All six provided some measure of appropriate diagnosis and treatment of depression. Screening most commonly occurred in the first 8 weeks postpartum; five of the six studies used the EPDS as the screening tool. Strength of evidence and findings are shown in Table E.

Table E. Strength-of-evidence domains for the effect of timing of screening on rates of referral and treatment among women with a positive screening test for postpartum depression						
		Domains Pertaining to SOE				
Timing	Number of Studies (Subjects)	Risk of Bias	Consistency	Directness	Precision	SOE and Magnitude of Effect (95% CI)
Prenatal vs. postpartum	3 (1,263)	Medium	Inconsistent	Direct	Imprecise	Low SOE for higher rates of referral/diagnosis prenatally
Delivery vs. postpartum	1 (230)	Low	NA	Direct	Imprecise	Low SOE for higher rates of referral/diagnosis during delivery admission

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

The main finding of these studies was that followup rates for women with positive screening tests were low, ranging from 0 to 30 percent, except in the fair-quality RCT, where screening, diagnosis, and treatment all occurred within the same practice setting. In one observational study, referral rates were significantly higher in women with abnormal screening test results during the delivery admission compared with 36 weeks gestation or 6 weeks postpartum.

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 the findings in 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.

Our findings are also consistent with the findings of the review conducted for two documents published in 2009, the United States Preventive Services Task Force (USPSTF) update for screening in adults[6] and the Institute of Medicine report on depression in parents,[5] both of which noted similar methodological issues in the literature as the 2005 AHRQ report did. Both reports also noted that there is reasonable evidence that screening for depression in adults can be effective if there are appropriate systems in place to assure that those with positive results are referred to appropriate diagnostic and therapeutic services; the USPSTF recommendations explicitly separate the recommendations based on the presence of such systems, with a "B" recommendation for screening if systems are in place but a "C" recommendation *against* screening without such systems.

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.[14]

Many included studies recruited populations whose demographics differed considerably from those of patients in the broader community. Overall, only 30 percent of included studies were conducted in the United States; the largest percentage was conducted in Europe or the UK (48 percent). Event rates for postpartum depression differ significantly between countries 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 nonresponse or attrition during the study period, thus limiting the applicability of the 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 sensitivity and specificity was substantially higher than point-prevalence estimates for the U.S. population, suggesting 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. CIs 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 period in a hospital setting before discharge; 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 directly compared 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 with screening, there are limited data on other maternal or infant health outcomes. Lastly, there is limited information on paternal outcomes.

The single U.S.-based study that demonstrated high rates of receipt of appropriate services and significant reductions with screening did so within the context of family physician practices where integrated screening, diagnosis, and treatment services were available. Because family physicians provide less than 10 percent of obstetric care and less than 20 percent of well-child visits in the United States, these results may not be directly applicable to the clinical settings that provide screening opportunities for most women in the first postpartum year.

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 to be performed, uncertainty about which instrument to use. These uncertainties are reflected in the recommendations by various stakeholder organizations discussed in the Introduction of our full CER. 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 USPSTF gives screening for depression in adults a "B" recommendation "when staff-assisted depression care supports are in place to assure accurate diagnosis, effective treatment, and follow-up" and a "C" recommendation against routine screening "when staff-assisted depression care supports are not in place."[6] Since the current evidence suggests that the prevalence of depression in postpartum women is similar overall to that in other women of reproductive age, these recommendations should be as applicable to women during the postpartum period as at any other time. Our evidence review found low rates of appropriate followup in the majority of studies, with a notable exception in a trial where screening, diagnosis, and treatment were all available within the same primary care setting,[15] which is consistent with the background review of screening for depression in the adult population conducted for the USPSTF.

If screening for depression during the postpartum period is especially important because of the potential impact on both mother and child, and if screening for depression is effective only when adequate resources are available to ensure appropriate followup, then the major policy implication of this report is that much greater attention needs to be paid to an explicit definition of the goals of a postpartum depression screening strategy. Our simulation results suggest that no matter what methods are used to ensure appropriate followup, the resources required are directly dependent on the test characteristics of the screening test. Table F shows the impact of test sensitivity and specificity and the prevalence of depression on the annual number of expected true positives, false positives, and false negatives from a one-time screen for postpartum depression when sensitivity and specificity are in the 80–90% range and inversely correlated (consistent with our review).

Table F. Effect of prevalence of major depression on annual expected true positives, false positives, and false negatives in the United States at varying levels of sensitivity and specificity assuming a one-time postpartum screen

Prevalence of Major Depression	Screening Results	Sensitivity 90%, Specificity 80%	Sensitivity 85%, Specificity 85%	Sensitivity 80%, Specificity 90%
4%	True positives	144,000	136,000	128,000
	False positives	768,000	576,000	384,000
	False negatives	16,000	24,000	32,000
8%	True positives	288,000	272,000	256,000
	False positives	736,000	552,000	368,000
	False negatives	32,000	48,000	64,000
15%	True positives	540,000	510,000	480,000
	False positives	680,000	510,000	340,000
	False negatives	60,000	90,000	120,000

This impact is magnified if women are screened multiple times during the postpartum period. Our modeling suggests that serial testing using a highly sensitive test (such as the "two-question screen") followed by the use of a more specific test results in substantial reductions in false positives with a much smaller increase in false negatives, and validation of this approach should be a high research priority. The choice of optimal test and test thresholds, testing algorithms, and test frequency need to be made based on an explicit consideration of the tradeoff between false-positive and false-negative results, including the necessity for adequate resources for managing women with positive screening results.

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 outcome 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 values of number of deliveries that are not integers have no interpretation) but does not reflect the underlying distribution.

For many of the recommendations below, formal simulation and decision models may prove useful. As described above, even a simple model can be helpful in illustrating tradeoffs and can highlight the relationship between uncertainty about the relative likelihood of adverse outcomes compared to favorable outcomes, the acceptable harm/benefit tradeoff, and the extent to which further research will help clarify the optimal decision or recommendation. This approach can be done using specific clinical outcomes only or explicitly incorporating costs; in the latter case, this value-of-information analysis can help inform research prioritization and research budgeting.[12,16] 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.

For all of the KQs, there is a general lack of evidence on the effectiveness of targeting fathers or both parents.

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 the following:

    1. Achieving consensus on the appropriate tradeoff between false positives and false negatives and using thresholds defined by these clinical criteria to determine optimal sensitivity and specificity for candidate screening instruments. As discussed above, even fairly small differences in test characteristics can translate into large differences in the likelihood of an accurate test result, with significant implications for both the individual patient and the larger health care system.

    2. Determining 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;[17] these data could then be used to inform the choice of which instruments to evaluate further.

    3. Defining 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. Directly comparing candidate instruments, either by having the same subject use each instrument

(randomized as to order of administration) or by randomizing different subjects to different instruments. The tradeoff here is between the increased generalizability of having subjects take a single test versus overall sample size.

5. Including an explicit discussion of screening frequency during the postpartum period, since this has significant implications for both the cumulative probability of a false-positive result as well as for the setting where screening is most likely to occur.

- The question of whether different instruments are better at identifying specific signs and symptoms is important only if there are effective interventions for those specific signs and symptoms. In order to discuss potential research designs, clarity is needed on which signs and symptoms are to be identified and what potential interventions are available. 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.
- If a large part of the goal of screening for depression is to improve longer term child outcome through improved functioning of the mother-infant dyad, then consideration should be given to characterizing the sensitivity and specificity of screening tests or algorithms, both existing ones and new ones, based on their ability to predict or detect maladaptive functioning or longer term adverse outcomes.

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, which is used to identify women at higher risk of breast cancer for more aggressive screening protocols). The potential impact of such a model could be estimated based on the absolute risk of postpartum depression at different thresholds and then using this information to estimate the number of false positives and false negatives resulting from screening only women identified as high risk. This estimate could be compared with 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

- There was insufficient direct evidence to address the effect of timing, setting, or provider on test characteristics. It seems plausible that differences in clinical outcomes relevant to timing, setting, or provider are more directly related to aspects of the process of screening, referral, and diagnosis than to differences in the test characteristics of the specific screening instrument used in the study. In other words, studies that compare the effects of timing, setting, or provider on overall clinical outcomes should be a higher priority for research resources than studies that only compare sensitivity and specificity of screening instruments by timing, setting, or provider.
- Additional RCTs comparing organized screening with usual care 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 and would necessarily be part of any formal discussion of appropriate targets for sensitivity and specificity.
- The use of a two-question screen followed by a standardized screening instrument in women who answer yes to one of the questions would appear to have substantial potential to improve screening efficiency based on reported test characteristics and a simple model; future screening studies in the United States should strongly consider including this approach as one of the study arms.
- Ideally, 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.
- If longer term studies are not feasible and the rationale for screening during the postpartum period is strengthened by the potential to improve longer term outcomes through improving the maternal-infant relationship, then studies should incorporate valid and sensitive measures of this relationship that are reliable surrogates for longer term outcomes. To the extent that scores on measures of depression may be more sensitive to depression treatment than scores on measures of parental function, consideration should be given to designing and powering studies to detect clinically meaningful differences in parental functioning as the primary outcome. A depression

screening and intervention study powered to detect a difference in a parental functioning outcome would be likely to have sufficient power to detect improvement in depression symptoms, whereas the converse may not be the case.

- There was low-strength evidence that timing might affect likelihood of receiving appropriate diagnostic and therapeutic services, and reported receipt of appropriate diagnostic and therapeutic services was much higher in two studies where screening, diagnosis, and treatment were available from the same provider.

Conclusions

The USPSTF recommends screening for depression in adults when adequate resources are available to ensure appropriate diagnostic and therapeutic services. The current evidence for women in the postpartum period is consistent with that recommendation. The prevalence of depression is similar to that observed in other women of the same age who are not pregnant or postpartum; the sensitivity and specificity of the available screening tests are similar; and although there is no direct evidence of variability in outcomes by setting, indirect comparisons across a small number of studies suggest that the receipt of appropriate services is much higher and depressive symptoms are substantially improved when screening, diagnosis, and treatment are provided by the same provider or practice. The ideal characteristics of a screening test for postpartum depression, including sensitivity, specificity, timing, and frequency, have not been defined. Because the balance of benefits and harms, at both the individual level and health system level, is highly dependent on these characteristics, broad consensus on these characteristics is needed.

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). AHRQ Publication No. 05-E006-2. Rockville, MD: Agency for Healthcare Research and Quality; February 2005.

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 the RTI-University of North Carolina Evidence-based Practice Center, under Contract No. 290-02-0016.) AHRQ Publication No. 05-E006-2. Rockville, MD: Agency for Healthcare Research and Quality. February 2005. www.ncbi.nlm.nih.gov/books/NBK37740/.

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. Institute of Medicine. Depression in Parents, Parenting, and Children: Opportunities to Improve Identification, Treatment, and Prevention. 2009. http://books.nap.edu/openbook.php?record_id=12565. Accessed October 9, 2012.

6. O'Connor EA, Whitlock EP, Gaynes B, et al. Screening for Depression in Adults and Older Adults in Primary Care: An Updated Systematic Review. Evidence Synthesis No. 75. AHRQ Publication No. 10-05143-EF-1. Rockville, Maryland: Agency for Healthcare Research and Quality, December 2009. www.ncbi.nlm.nih.gov/books/NBK36406/. Accessed October 10, 2012. PMID: 20722174.

7. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. Chapters available at www.effectivehealthcare.ahrq.gov. Accessed January 3, 2012.

8. Agency for Healthcare Research and Quality. Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC017. Rockville, MD: Agency for Healthcare Research and Quality. www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=558. Accessed January 3, 2012.

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

10. 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.

11. 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.

12. 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. www.effectivehealthcare.ahrq.gov. Accessed January 3, 2012.

13. 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.

14. 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.

15. Yawn BP, Dietrich AJ, Wollan P, et al. TRIPPD: a practice-based network effectiveness study of postpartum depression screening and management. Ann Fam Med. 2012;10(4):320-9. PMID: 22778120.

16. 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. www.pcori.org/assets/Value-of-Information-Analysis-for-Patient-Centered-Outcomes-Research-Prioritization.pdf. Accessed July 3, 2012.

17. 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.

Full Report

This executive summary is part of the following document: Myers ER, Aubuchon-Endsley N, Bastian LA, Gierisch JM, Kemper AR, Swamy GK, Wald MF, McBroom AJ, Lallinger KR, Gray RN, Green C, Sanders GD. Efficacy and Safety of Screening for Postpartum Depression. Comparative Effectiveness Review 106. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-2007-10066-I.) AHRQ Publication No. 13-EHC064-EF. Rockville, MD: Agency for Healthcare Research and Quality; April 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

AHRQ Pub. No. 13-EHC064-1-EF
April 2013