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

Contract No. 290-2007-10066-I

Prepared by:
Duke Evidence-based Practice Center
Durham, NC

Investigators:
Evan R. Myers, M.D., M.P.H.
Nicki Aubuchon-Endsley, Ph.D.
Lori A. Bastian, M.D.
Jennifer M. Gierisch, Ph.D., M.P.H.
Alex R. Kemper, M.D., M.P.H., M.S.
Geeta K. Swamy, M.D.
Marla F. Wald, M.D.
Amanda J. McBroom, Ph.D.
Kathryn R. Lallinger, M.S.L.S.
Rebecca N. Gray, D.Phil.
Cindy Green, Ph.D.
Gillian D. Sanders, Ph.D.

AHRQ Publication No. 13-EHC064-EF
April 2013
This report is based on research conducted by the Duke Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10066-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

This document is in the public domain and may be used and reprinted without special permission. Citation of the source is appreciated.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of systematic reviews to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. These reviews provide comprehensive, science-based information on common, costly medical conditions, and new health care technologies and strategies.

Systematic reviews are the building blocks underlying evidence-based practice; they focus attention on the strength and limits of evidence from research studies about the effectiveness and safety of a clinical intervention. In the context of developing recommendations for practice, systematic reviews can help clarify whether assertions about the value of the intervention are based on strong evidence from clinical studies. For more information about AHRQ EPC systematic reviews, see www.effectivehealthcare.ahrq.gov/reference/purpose.cfm.

AHRQ expects that these systematic reviews will be helpful to health plans, providers, purchasers, government programs, and the health care system as a whole. Transparency and stakeholder input are essential to the Effective Health Care Program. Please visit the Web site (www.effectivehealthcare.ahrq.gov) to see draft research questions and reports or to join an email list to learn about new program products and opportunities for input.

We welcome comments on this systematic review. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Suchitra Iyer, Ph.D.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality
Acknowledgments

The authors thank Megan von Isenburg, M.S.L.S., for help with the literature search and retrieval; Rachael Posey, M.S.L.S., for assistance with project coordination; and William Meyer, M.S.W., L.C.S.W., for assistance during the topic refinement phase.

Key Informants

In designing the study questions, the EPC consulted several Key Informants who represent the end-users of research. The EPC sought the Key Informant input on the priority areas for research and synthesis. Key Informants are not involved in the analysis of the evidence or the writing of the report. Therefore, in the end, study questions, design, methodological approaches, and/or conclusions do not necessarily represent the views of individual Key Informants.

Key Informants must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their role as end-users, individuals with potential conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any conflicts of interest.

The list of Key Informants who participated in developing this report follows:

Lisa Begg, Dr.P.H., R.N.
Director of Research Programs, Office of Research on Women’s Health
National Institutes of Health
Bethesda, MD

Wendy N. Davis, Ph.D.
Executive Director, Postpartum Support International
Portland, OR

Denise M. Dougherty, Ph.D.
Senior Advisor, Child Health and Quality Improvement
Agency for Healthcare Research and Quality
Rockville, MD

Marian Earls, M.D.
Former Medical Director, Guilford Child Health, Inc.
Member, Committee on the Psychosocial Aspects of Child and Family Health,
American Academy of Pediatrics
Greensboro, NC

Tina Groat, M.D., M.B.A.
National Medical Director for Women’s Health, United Healthcare
Edina, MN

Karen Hench, M.S., R.N.
Past Acting Division Director, Division of Healthy Start and Perinatal Services
Maternal and Child Health Bureau
Rockville, MD

Deborah Karsnitz, D.N.P., C.N.M.
Faculty Course Coordinator, Frontier Nursing University
President, Kentucky State Affiliate, American College of Nurse Midwives
Louisville, KY

Susan G. Kornstein, M.D.
Immediate Past President, International Association for Women’s Mental Health
Professor of Psychiatry and Obstetrics/Gynecology, Virginia Commonwealth University
Richmond, VA
Technical Expert Panel

In designing the study questions and methodology at the outset of this report, the EPC consulted several technical and content experts. Broad expertise and perspectives were sought. Divergent and conflicted opinions are common and perceived as healthy scientific discourse that results in a thoughtful, relevant systematic review. Therefore, in the end, study questions, design, methodological approaches, and/or conclusions do not necessarily represent the views of individual technical and content experts.

Technical Experts must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential conflicts of interest identified.

The list of Technical Experts who participated in developing this report follows:

William H. Barth, Jr., M.D.
Associate Professor of Obstetrics,
Gynecology and Reproductive Biology,
Harvard Medical School
Chair, Committee on Obstetric Practice,
American College of Obstetricians and
Gynecologists
Boston, MA

Lisa Begg, Dr.P.H., R.N.
Director of Research Programs, Office of
Research on Women’s Health
National Institutes of Health
Bethesda, MD

Wendy N. Davis, Ph.D.
Executive Director, Postpartum Support
International
Portland, OR

Denise M. Dougherty, Ph.D.
Senior Advisor, Child Health and Quality
Improvement
Agency for Healthcare Research and Quality
Rockville, MD
Marian Earls, M.D.
Former Medical Director, Guilford Child Health, Inc.
Member, Committee on the Psychosocial Aspects of Child and Family Health, American Academy of Pediatrics
Greensboro, NC

Tina Groat, M.D., M.B.A.
National Medical Director for Women’s Health, United Healthcare
Edina, MN

Karen Hench, M.S., R.N.
Past Acting Division Director, Division of Healthy Start and Perinatal Services
Maternal and Child Health Bureau
Rockville, MD

Deborah Karsnitz, D.N.P., C.N.M.
Faculty Course Coordinator, Frontier Nursing University
President, Kentucky State Affiliate, American College of Nurse Midwives
Louisville, KY

Susan G. Kornstein, M.D.
Immediate Past President, International Association for Women’s Mental Health
Professor of Psychiatry and Obstetrics/Gynecology, Virginia Commonwealth University
Richmond, VA

Virginia Moyer, M.D., M.P.H.
Professor of Pediatrics, Baylor College of Medicine
Chair, United States Preventive Services Task Force
Houston, TX

Michael O’Hara, Ph.D.
Professor of Psychology and Starch Faculty Fellow
University of Iowa
Iowa City, IA

Peter J. Schmidt, M.D.
Chief, Section on Behavioral Endocrinology
National Institute of Mental Health
Bethesda, MD

Caroline Signore, M.D., M.P.H.
Program Scientist, Pregnancy and Perinatal Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development
Fellow, American College of Obstetricians and Gynecologists
Rockville, MD

Barbara Yawn, M.D., M.Sc.
Director of Research, Olmsted Medical Center
Adjunct Professor, Department of Family and Community Health, University of Minnesota
Rochester, MN

Kimberly Ann Yonkers, M.D.
Professor of Psychiatry and Obstetrics, Gynecology, and Reproductive Sciences, Yale School of Medicine
New Haven, CT
Peer Reviewers

Prior to publication of the final evidence report, the EPC sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual Peer Reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential nonfinancial conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential nonfinancial conflicts of interest identified.

The list of Peer Reviewers follows:

Lee Cohen, M.D.
Director, Perinatal and Reproductive Psychiatry Clinical Research Program,
Massachusetts General Hospital
Associate Professor of Psychiatry, Harvard Medical School
Boston, MA

Emily Feinberg, Sc.D., M.Sc., M.S.N.
Associate Professor of Pediatrics, Boston University School of Medicine
Boston, MA

Patricia Fontaine, M.D., M.S.
Associate Professor, Department of Family Medicine and Community Health
University of Minnesota
Minneapolis, MN

Katherine Hartmann, M.D., Ph.D.
Professor of Medicine and of Obstetrics and Gynecology, Vanderbilt University School of Medicine
Deputy Director, Institute of Medicine and Public Health
Nashville, TN

Jane Honikman, M.S.
Founder and Former Executive Director,
Postpartum Support International
Santa Barbara, CA

Samantha Meltzer-Brody, M.D.
Associate Professor and Director, Perinatal Psychiatry Program
University of North Carolina Center for Women’s Mood Disorders
Chapel Hill, NC

Sarah Hudson Scholle, Dr.PH., M.P.H.
Vice President of Research and Analysis,
National Committee for Quality Assurance
Washington, DC
Efficacy and Safety of Screening for Postpartum Depression

Structured Abstract

Objectives. To describe the benefits and harms of specific 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, to July 24, 2012, that evaluated the performance of screening instruments for postpartum depression, potential benefits and harms of screening, and impact on appropriate postscreening 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. Forty studies (represented by 45 articles) were identified as relevant to this review. Eighteen studies provided sensitivity and specificity data on 9 screening instruments: 11 on the Edinburgh Postnatal Depression Scale, 4 on the Postpartum Depression Screening Scale, 4 on different versions of the Beck Depression Inventory, 2 on a “two-question” screen, and 1 each on 5 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; the two-question screen had 100 percent sensitivity but specificities of 45–65 percent. Fifteen studies analyzed the association between risk factors and postpartum depression. Although adverse pregnancy outcomes and chronic medical conditions (low strength of evidence) and past history of depression, poor relationship quality, and poor social support (moderate strength of evidence) were all associated with an increased risk of postpartum depression, only two studies directly reported an effect on test results. (Sensitivity was nonsignificantly increased in primigravidas compared with multigravidas.) 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 five studies, there was low to moderate strength of evidence that screening resulted in decreased depressive symptoms and improved mental health; in four of these studies, improvement in depressive symptoms was not accompanied by improvement in measures of parenting stress. Rates of referral and treatment for women with positive screening results were substantially higher in two studies where screening, diagnosis, and treatment were provided in the same setting; referral rates in other studies were all 50 percent or less. Modeling suggests that serial testing with a two-question screen followed by a second more specific instrument for those who have a positive result may be a reasonable strategy to reduce false positives while minimizing false negatives.
**Conclusions.** The potential effectiveness of screening for postpartum depression appears to be related to the availability of systems to ensure adequate followup of women with positive results. 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.
Contents

Executive Summary ........................................................................................................................................ ES-1
Introduction .................................................................................................................................................. 1
  Background ............................................................................................................................................. 1
  Postpartum Depression .......................................................................................................................... 1
  Adverse Outcomes Associated With Postpartum Depression ............................................................ 2
  Screening for Postpartum Depression ................................................................................................. 2
  Clinical and Socioeconomic Factors Affecting Risk for Postpartum Depression .............................. 5
Scope and Key Questions .......................................................................................................................... 7
  Scope of the Review .............................................................................................................................. 7
  Key Questions ....................................................................................................................................... 9
Analytic Framework ................................................................................................................................... 10
Methods .................................................................................................................................................... 12
  Topic Refinement and Review Protocol .............................................................................................. 12
  Literature Search Strategy .................................................................................................................. 12
    Search Strategy ................................................................................................................................... 12
    Inclusion and Exclusion Criteria ......................................................................................................... 13
    Study Selection ..................................................................................................................................... 16
    Data Extraction .................................................................................................................................... 17
  Quality Assessment of Individual Studies ............................................................................................ 17
  Data Synthesis ....................................................................................................................................... 18
  Strength of the Body of Evidence ........................................................................................................ 20
  Applicability ......................................................................................................................................... 21
  Peer Review and Public Commentary .................................................................................................. 21
Results ......................................................................................................................................................... 22
  Introduction .......................................................................................................................................... 22
  Results of Literature Searches ............................................................................................................. 22
  Description of Included Studies .......................................................................................................... 22
  Key Question 1. Performance Characteristics of Screening Instruments ........................................... 24
    Key Points .......................................................................................................................................... 24
    Description of Included Studies ........................................................................................................ 25
    Detailed Synthesis ............................................................................................................................. 26
  Key Question 2. Effect of Individual Subject Factors on Screening Performance .............................. 32
    Key Points .......................................................................................................................................... 32
    Description of Included Studies ........................................................................................................ 33
    Detailed Synthesis ............................................................................................................................. 33
  Key Question 3. Effect of Testing Variables on Screening Performance ........................................... 41
    Key Points .......................................................................................................................................... 41
    Description of Included Studies ........................................................................................................ 41
    Detailed Synthesis ............................................................................................................................. 41
  Key Question 4. Comparative Benefits of Screening .......................................................................... 42
    Key Points .......................................................................................................................................... 43
    Description of Included Studies ........................................................................................................ 43
    Detailed Synthesis ............................................................................................................................. 43
  Key Question 5. Comparative Harms of Screening .............................................................................. 46

x
Key Points ......................................................................................................................... 46
Description of Included Study .......................................................................................... 46
Detailed Synthesis ............................................................................................................. 46
Key Question 6. Factors Affecting the Likelihood of an Appropriate Action After a Positive Screening Result ................................................................. 47
Key Points ......................................................................................................................... 47
Description of Included Studies ........................................................................................ 47
Detailed Synthesis ............................................................................................................. 48
Discussion ................................................................................................................................... 52
Key Findings and Strength of Evidence ................................................................................. 52
KQ 1. Performance Characteristics of Screening Instruments ......................................... 52
KQ 2. Effect of Individual Factors on Screening Performance ........................................ 56
KQ 3. Effect of Testing Variables on Screening Performance ......................................... 59
KQ 4. Comparative Benefits of Screening and KQ 5. Comparative Harms of Screening ........................................................................................................ 59
KQ 6. Factors Affecting the Likelihood of an Appropriate Action After a Positive Screening Result ......................................................................................... 61
Findings in Relation to What Is Already Known .............................................................. 61
Applicability ................................................................................................................................... 63
Implications for Clinical and Policy Decisionmaking ............................................................ 64
Potential Value of Simulation Modeling ............................................................................. 65
Limitations of the Comparative Effectiveness Review Process ........................................ 69
Limitations of the Evidence Base ........................................................................................... 70
Research Gaps ..................................................................................................................... 71
General Gaps ..................................................................................................................... 71
KQ 1 .................................................................................................................................. 72
KQ 2 .................................................................................................................................. 73
KQs 3–6 ............................................................................................................................ 73
Conclusions ................................................................................................................................... 75
References .......................................................................................................................... 76
Abbreviations ....................................................................................................................... 84

Tables
Table A. Strength-of-evidence domains for test characteristics of screening tests for postpartum depression ........................................................................................................ ES-9
Table B. Strength-of-evidence domains for associations with patient characteristics and risk of postpartum depression ..................................................................................... ES-12
Table C. Strength-of-evidence domains for the effect of varying timing on screening for postpartum depression .................................................................................................... ES-14
Table D. Strength-of-evidence domains for benefits and harms of screening for postpartum depression ............................................................................................................. ES-15
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 ..................................................................................... ES-16
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............................. ES-18
Table 1. DSM-IV-TR diagnostic criteria for major depressive disorder........................................ 1
Table 2. Guidelines/recommendations for screening for postpartum depression............................ 7
Table 3. Inclusion and exclusion criteria...................................................................................... 13
Table 4. Definitions of overall quality ratings............................................................................... 18
Table 5. Strength of evidence—required domains ....................................................................... 20
Table 6. Test characteristics for postpartum depression screening instruments other than EPDS or PDSS................................................................................................................ 29
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.......................................... 38
Table 12. Relationship and social support risk factors for postpartum major depression............ 39
Table 13. Sensitivity and specificity of screening instruments in the early and late postpartum periods .......................................................................................................................... 42
Table 14. Strength-of-evidence domains for test characteristics of screening tests for postpartum depression ......................................................................................................... 53
Table 15. 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 ........................................................... 56
Table 16. Strength-of-evidence domains for associations with patient characteristics and risk of postpartum depression .................................................................................................. 57
Table 17. Strength-of-evidence domains for the effect of varying timing on screening for postpartum depression ............................................................................................................ 59
Table 18. Strength-of-evidence domains for benefits and harms of screening for postpartum depression .......................................................................................................................... 60
Table 19. Strength-of-evidence domains for the effect of timing on rates of referral and treatment among women with a positive screening test for postpartum depression........ 61
Table 20. Estimated annual number of true positives, false positives, and false negatives in the United States from screening with “single test” versus “serial tests” 66
Table 21. Estimated number of true positives, false positives, and false negatives with screening at postpartum and well-child visits ............................................................... 69

Figures
Figure 1. Analytic framework ....................................................................................................... 10
Figure 2. Literature flow diagram ................................................................................................. 23
Figure 3. Sensitivity of the EPDS at various thresholds............................................................... 26
Figure 4. Specificity of the EPDS at various thresholds............................................................... 27
Figure 5. Sensitivity of the PDSS at various thresholds............................................................... 28
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
Appendixes
Appendix A. Exact Search Strings
Appendix B. Data Abstraction Elements
Appendix C. Included Studies
Appendix D. Excluded Studies
Appendix E. Study Characteristics Table
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 percent of 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

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.\textsuperscript{3,5}

Depression in adults has a significant impact on quality of life, productivity, and social functioning,\textsuperscript{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,\textsuperscript{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 and Medical Test Guide. 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) 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. 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. 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 and Medical Test Guide. 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 and the Medical Test Guide. 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

<table>
<thead>
<tr>
<th>Screening Test</th>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Test Performance (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antenatal Risk Questionnaire</td>
<td>Sensitivity</td>
<td>1 (276)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (276)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td>BDI</td>
<td>Sensitivity</td>
<td>2 (1,151)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>2 (1,151)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>BDI-II</td>
<td>Sensitivity</td>
<td>2 (650)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>2 (650)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Screening Tests</td>
<td>Outcome</td>
<td>Number of Studies (Subjects)</td>
<td>Risk of Bias</td>
<td>Consistency</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------</td>
<td>-----------------------------</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>EPDS</strong></td>
<td>Sensitivity</td>
<td>11 (3,456)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>11 (3,456)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td><strong>HRSD-17</strong></td>
<td>Sensitivity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td><strong>HRSD-21</strong></td>
<td>Sensitivity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Leverton Questionnaire</strong></td>
<td>Sensitivity</td>
<td>1 (617)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (617)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td><strong>PDSS</strong></td>
<td>Sensitivity</td>
<td>4 (903)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>4 (903)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
</tbody>
</table>
Table A. Strength-of-evidence domains for test characteristics of screening tests for postpartum depression (continued)

<table>
<thead>
<tr>
<th>Screening Tests</th>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Test Performance (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>Sensitivity</td>
<td>1 (506)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (506)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td>Two-Question Screen</td>
<td>Sensitivity</td>
<td>2 (600)</td>
<td>Low</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>2 (600)</td>
<td>Low</td>
<td>Consistent</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Risk Factor</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
<td>Directness</td>
</tr>
<tr>
<td>Maternal Demographics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>3 (5,578)</td>
<td>Medium</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Education</td>
<td>2 (4,757)</td>
<td>Medium</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Income</td>
<td>1 (4,245)</td>
<td>Medium</td>
<td>NA</td>
</tr>
<tr>
<td>Employment status (unemployed vs. employed)</td>
<td>1 (363)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td>Obstetric History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parity</td>
<td>2 (4,998)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Preterm/low birthweight infant</td>
<td>2 (4,711)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Smoking</td>
<td>2 (4,998)</td>
<td>Medium</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>1 (4,348)</td>
<td>Medium</td>
<td>NA</td>
</tr>
<tr>
<td>General Medical History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor health status/chronic illness</td>
<td>2 (4,993)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Obesity</td>
<td>1 (598)</td>
<td>Medium</td>
<td>NA</td>
</tr>
<tr>
<td>Psychiatric History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of perinatal depression</td>
<td>2 (1,082)</td>
<td>High</td>
<td>Consistent</td>
</tr>
<tr>
<td>History of depression</td>
<td>5 (2,057)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>History of premenstrual dysphoric disorder</td>
<td>1 (210)</td>
<td>Medium</td>
<td>NA</td>
</tr>
<tr>
<td>Any psychiatric diagnosis</td>
<td>2 (1,075)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Anxiety</td>
<td>2 (1,305)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Personality (vulnerable/ neuroticism)</td>
<td>2 (685)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
</tbody>
</table>
Table B. Strength-of-evidence domains for associations with patient characteristics and risk of postpartum depression (continued)

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
<td>Directness</td>
</tr>
<tr>
<td>Relationship/ Social Support</td>
<td>Marital status (single/no relationship)</td>
<td>3 (5,803)</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Poor relationship quality</td>
<td>5 (6,101)</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Poor social support</td>
<td>4 (1,830)</td>
<td>Medium</td>
</tr>
</tbody>
</table>

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).

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

<table>
<thead>
<tr>
<th>Timing</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery to 8 weeks vs. 8 weeks to 6 months</td>
<td>1 (534)</td>
<td>Risk of Bias</td>
<td>High</td>
</tr>
<tr>
<td>Delivery vs. 6 weeks</td>
<td>1 (113)</td>
<td>Consistency</td>
<td>NA</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Benefits/ Harms</th>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Risk of Bias</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefits</td>
<td>Depressive symptoms</td>
<td>5 (8,071)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
</tr>
<tr>
<td>Benefits</td>
<td>Mental health score (SF-12)</td>
<td>1 (2,579)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
</tr>
<tr>
<td>Benefits</td>
<td>Parental stress</td>
<td>4 (5,567)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
</tr>
<tr>
<td>Harms</td>
<td>Unscheduled doctor visits for infant</td>
<td>1 (462)</td>
<td>Medium</td>
<td>NA</td>
<td>Direct</td>
</tr>
</tbody>
</table>

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

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 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.

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.
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

<table>
<thead>
<tr>
<th>Timing</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
</tr>
<tr>
<td>Prenatal vs. postpartum</td>
<td>3 (1,263)</td>
<td>Medium</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Delivery vs. postpartum</td>
<td>1 (230)</td>
<td>Low</td>
<td>NA</td>
</tr>
</tbody>
</table>

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. 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 and the Institute of Medicine report on depression in parents, 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.

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.” 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, 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**

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

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. 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; 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


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 (Note: 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 of postpartum depression to extend beyond 4 weeks and up to 12 months after delivery, and to include a “minor depression” subcategory (2 to 4 of the symptoms listed in Table 1).

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

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.</td>
<td>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.)</td>
</tr>
<tr>
<td></td>
<td>• 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)</td>
</tr>
<tr>
<td></td>
<td>• 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)</td>
</tr>
<tr>
<td></td>
<td>• 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</td>
</tr>
<tr>
<td></td>
<td>• Insomnia or hypersomnia nearly every day</td>
</tr>
<tr>
<td></td>
<td>• Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down)</td>
</tr>
<tr>
<td></td>
<td>• Fatigue or loss of energy nearly every day</td>
</tr>
<tr>
<td></td>
<td>• Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick)</td>
</tr>
<tr>
<td></td>
<td>• Diminished ability to think or concentrate, or indecisiveness, nearly every day (either subjective account or as observed by others)</td>
</tr>
<tr>
<td></td>
<td>• 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</td>
</tr>
<tr>
<td>B.</td>
<td>The symptoms do not meet the criteria for mixed episode (DSM-IV-TR, p.365).</td>
</tr>
<tr>
<td>C.</td>
<td>The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.</td>
</tr>
<tr>
<td>D.</td>
<td>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).</td>
</tr>
<tr>
<td>E.</td>
<td>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.</td>
</tr>
</tbody>
</table>

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

The most recent U.S.–based formal synthesis of the evidence, performed for the Agency for Healthcare Research and Quality (AHRQ) in 2005,²³ 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% CI, 15.1 to 30.0%). 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 nonpregnant/nonpostpartum women of reproductive age. However, the prevalence of depressive symptoms not meeting diagnostic criteria for depression may be higher, particularly in the first 3 months after birth. Consistent limitations in the literature noted by the AHRQ review included small sample size (precluding subgroup analyses) and lack of generalizability.

Adverse Outcomes Associated With Postpartum Depression

Depression in adults has a significant impact on quality of life, productivity, and social functioning, and there is no evidence that these effects are any different for women during the postpartum period. In addition to the substantial morbidity burden, there is an appreciable risk of mortality through suicide. Although the risk of suicide in women may be lower during pregnancy and the postpartum period, a review of maternal mortality in the United Kingdom (UK) during the 1990s found that suicide was the leading cause of maternal mortality, accounting for 29 percent of maternal deaths. In addition, postpartum depression may increase the risk of infant mortality through neglect, abuse, or homicide.

The impact of depression on mothers alone is sufficient to justify intervention, although the available evidence suggests that the burden of disease, particularly for major depression, is similar during the postpartum period and other times not associated with pregnancy or recent childbirth. However, there is another rationale for giving special consideration to the effective prevention or treatment of depression during the postpartum period. In 2009, the Institute of Medicine (IOM) published a comprehensive review of the impact of depression in parents on both parental and child outcomes. In observational studies, maternal depression is consistently associated with adverse effects on maternal–infant interactions and some measures of infant development, and this evidence has been explicitly cited as part of the rationale for screening by providers who care for infants. For example, the IOM report found an increased risk of “maladaptive” utilization, including underutilization of primary and preventive services, increased use of emergency services, and increased risk of hospitalization even after adjusting for infant health status.

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 to improve outcomes for both mother and child. 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

There is high-quality evidence that effective treatments are available for patients who meet criteria for major depression in other settings, and the available evidence suggests that both pharmacological\(^1\) and nonpharmacological\(^4,14\) treatments can be effective in the postpartum setting. Given the availability of effective treatment, screening instruments with acceptable test characteristics, and reliable systems for ensuring that women identified through a screening program receive appropriate diagnostic and therapeutic services, screening during the postpartum period is at least as justifiable as screening during other times. And, as noted below, screening for depression receives a “B” recommendation from the U.S. Preventive Services Task Force (USPSTF) if systems for ensuring receipt of appropriate services are in place.

Because maternal depression is consistently associated with adverse effects on maternal–infant interactions and longer term development outcomes, screening would be even more important if there were direct evidence that screening and treatment of previously undiagnosed maternal depression leads to improvement in these outcomes. However, 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 important outcomes—such as measures of infant health and development—have not been included,\(^2,3,15\) a deficiency noted in the IOM report.\(^4\) There is some evidence that treating depression in mothers improves some measures of child mental health and functioning,\(^16,17\) but these studies have not specifically been in the context of depression detected through screening in the postpartum period. Given the consistent association between depressive symptoms and adverse effects on maternal–infant interaction, it is possible that interventions performed in response to depressive symptoms not meeting the criteria for major depression could result in improved outcomes for the mother–infant dyad.

Potential Harms of Screening

In their 2009 recommendations on screening for depression in adults, the USPSTF 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 none of the reviewed studies provided any evidence regarding these potential harms.\(^18\) 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 a false-positive result 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 maternal outcomes.\(^2,3,19\) If a diagnosis of minor depression does not lead to effective treatment, then patients may be exposed to the potential side effects of therapy (particularly medical therapy) in addition to being labeled as depressed without a concomitant improvement in outcomes for themselves or their child.

Screening also requires resources. Even the use of self-administered tests requires some provider time to review and document the results, with additional time required for further evaluation of positive results, whether this evaluation takes place within the same setting as screening, or through referral. Even if screening leads to improved outcomes for depressed mothers or their infants, the resources required to screen all postpartum women and to evaluate
women with positive results (including women with false positive results) can place a burden on health care delivery systems, particularly public health systems, which may already be having difficulty meeting patient needs.

**Accuracy of Screening Instruments**

In evaluating strategies that involve 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 previous reviews have found them 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. In the context of screening for postpartum depression, higher sensitivity means more women with undiagnosed and untreated depression are detected, while higher specificity means fewer nondepressed women will need further evaluation to rule out depression. One advantage of reporting test sensitivity and specificity is that, because they are inherent characteristics of the tests themselves, sensitivity and specificity estimates for 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 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 in (1) maximizing detection of undiagnosed and untreated depression and (2) minimizing false-positive results. For example, the committee opinion on screening for depression during and after pregnancy developed by the American Congress of Obstetricians and Gynecologists 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 review on postpartum depression\textsuperscript{2,3} found that the sensitivity of all instruments reviewed was greater for a diagnosis of major depression alone compared with a broader definition of major or minor depression. As noted above, there is greater uncertainty about the availability of effective treatments for minor depression. This may be even more important in the setting of postpartum depression—if depressive symptoms that do not meet diagnostic criteria are associated with adverse effects on the mother–infant dyad, and treatment of these symptoms leads to improved developmental outcomes, then the “optimal” sensitivity target for a screening strategy should be based on these considerations.

**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.\textsuperscript{4,21-25} Because the outcomes of screening for any condition are dependent on the likelihood of that condition at the time of screening (the prior probability of disease), 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 physicians, 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, but issues
regarding scope of practice, legal liability, and appropriate referral remain challenges.

**Current Screening Recommendations**

All major organizations providing care to pregnant and postpartum women and infants
recognize the risks associated with 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 cited by these
organizations that limit the strength of recommendations include the lack of sufficient data on the
most appropriate screening instrument (including culturally appropriate tools to reflect
population diversity), 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.

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.
Table 2. Guidelines/recommendations for screening for postpartum depression

<table>
<thead>
<tr>
<th>Organization</th>
<th>Statement</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. Preventive Services Task Force&lt;sup&gt;18&lt;/sup&gt;</td>
<td>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.</td>
<td>December 2009</td>
</tr>
<tr>
<td>American College of Obstetricians and Gynecologists, Committee on Obstetric Practice&lt;sup&gt;20&lt;/sup&gt;</td>
<td>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.</td>
<td>February 2010</td>
</tr>
<tr>
<td>American Academy of Pediatrics, Committee on Psychosocial Aspects of Child and Family Health&lt;sup&gt;26&lt;/sup&gt;</td>
<td>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.</td>
<td>November 2010</td>
</tr>
<tr>
<td>American Academy of Family Physicians&lt;sup&gt;29&lt;/sup&gt;</td>
<td>No specific recommendations for postpartum depression; general recommendations for screening follow those of the U.S. Preventive Services Task Force.&lt;sup&gt;18&lt;/sup&gt;</td>
<td>October 2010</td>
</tr>
<tr>
<td>American College of Nurse Midwives&lt;sup&gt;30&lt;/sup&gt;</td>
<td>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.</td>
<td>December 2003</td>
</tr>
<tr>
<td>United Kingdom National Institute for Health and Clinical Excellence&lt;sup&gt;31&lt;/sup&gt;</td>
<td>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), health care 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.</td>
<td>April 2007</td>
</tr>
</tbody>
</table>

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

Scope and Key Questions

Scope of the Review

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. Further details are provided under “Key Questions” and “Analytic Framework,” below, and in the section on “Inclusion and Exclusion Criteria” in the
Methods chapter. As specified in the Key Questions (KQs), 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 a separate AHRQ CER (currently in progress) that will be published as a separate report.

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 focuses on the first two aspects of screening benefits; a separate evidence review of the effectiveness of treatment for perinatal depression is currently ongoing. 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 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 systemic level, and suggest areas where additional research to specifically addresses the limitations of the currently available evidence would help resolve this uncertainty.

**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 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 to with 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.

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. (Fathers and domestic partners were also considered, as specified in Table 3. For clarity, those groups are not depicted here.) 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 (such as proportion initiating breastfeeding or duration of 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. Note that 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 a separate AHRQ CER (currently in progress) that will be published as a separate report.
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)\(^32\) and the “Methods Guide for Medical Test Reviews” (hereafter referred to as the Medical Test Guide).\(^33\) 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.\(^34\)

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 and 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.\(^35\)

Literature Search Strategy

Search Strategy

To identify relevant published literature, we searched PubMed\(^\circ\), Embase\(^\circ\), PsycINFO\(^\circ\), 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 review on postpartum depression).\(^2,3\) Given the findings of the 2005 review regarding the level of evidence, we chose these dates after consultation with AHRQ, Key Informants, and the TEP in order to maximize efficiency. The primary impediment to formal data synthesis in the 2005 review was study heterogeneity. Therefore, it was unlikely that we would be able to combine literature identified in that report with newer data in any subsequent meta-analyses. This led us to conclude that qualitative comparison of our findings to those of prior reviews would be a more useful approach. Where possible, we used existing validated search filters (such as the Clinical Queries Filters in PubMed). An experienced search librarian guided all searches. Search dates and exact search strings are provided in Appendix A. We supplemented the electronic searches with a manual search of citations from a set of key primary and systematic review articles.\(^3,12,14,15,23,24,36-72\) The reference lists from these articles were hand-searched and cross-referenced against our library of database search results. Additional relevan
t articles not already under consideration were retrieved for screening. 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 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 dates and exact search terms used for these sources are provided in Appendix A. The search of ClinicalTrials.gov was also used as a mechanism to ascertain publication bias by identifying 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

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>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Populations</strong></td>
<td>• Pregnant women and women up to 12 months postpartum</td>
<td>• Women currently undergoing treatment for depression</td>
</tr>
<tr>
<td></td>
<td>• Subgroups of potential interest include:</td>
<td>• Studies where the primary objective is to detect depression during pregnancy rather than to identify risk factors for postpartum depression (studies that assessed women prenatally for risk of postpartum depression were not excluded)</td>
</tr>
<tr>
<td></td>
<td>o Race/ethnicity</td>
<td>• 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</td>
</tr>
<tr>
<td></td>
<td>o Income</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Parity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Cultural norms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o History of mood disorders</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Perinatal outcomes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o History of intimate partner violence</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3. Inclusion and exclusion criteria (continued)

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
</table>
| **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-1A, 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  
  o Patient Health Questionnaire (PRIME-MD PHQ) | • Validation studies, or screening conducted using a nonvalidated 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 nonvalidated 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 (e.g., proportion initiating breastfeeding or duration of breastfeeding)  
  o KQ 6:  
    ▪ Receipt of appropriate diagnostic and treatment services for symptoms of depression | • Outcomes measured predelivery  
• 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 |
Table 3. Inclusion and exclusion criteria (continued)

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcomes (continued)</td>
<td>• Final outcomes (KQ 4):</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Scores on validated diagnostic instruments for depression</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Health-related quality of life, based on validated measures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Maternal suicidal/infanticidal behaviors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Scores on validated instruments of infant health and development</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Maternal health system resource utilization, including number of visits and estimates of total and attributable costs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Infant health system resource utilization, including number of visits and estimates of total and attributable costs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>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)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Adverse effects (KQ 5):</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Scores on validated measures of stigmatization</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Health-related quality of life, based on validated measures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Maternal health system resource utilization, including number of visits and estimates of total and attributable costs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Infant health system resource utilization, including number of visits and estimates of total and attributable costs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>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)</td>
<td></td>
</tr>
<tr>
<td>Timing</td>
<td>• Intervention</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Prenatal period</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Immediate postpartum period (up to 6 weeks after delivery)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Up to 12 months after delivery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Followup</td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Begins at delivery; timing of followup not limited</td>
<td></td>
</tr>
<tr>
<td>Setting</td>
<td>• Any clinical provider setting, home</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Study locations include at least one high-income economy as defined by the World Bank,(^{73}) 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.</td>
<td>None</td>
</tr>
</tbody>
</table>

\(^{73}\) 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.
Table 3. Inclusion and exclusion criteria (continued)

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>• Original data</td>
<td>• Editorials, nonsystematic reviews, letters, case series, case reports</td>
</tr>
<tr>
<td></td>
<td>• 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</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• RCTs: All sample sizes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Observational studies: sample size ≥100 subjects</td>
<td></td>
</tr>
<tr>
<td>Publications</td>
<td>• English-language only</td>
<td>• Non-English-language articles&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>• Peer-reviewed articles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Relevant systematic review, meta-analysis, or methods article (to be used for background only)&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Published on or after January 1, 2004</td>
<td></td>
</tr>
</tbody>
</table>

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

<sup>a</sup>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.

<sup>b</sup>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.

### 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. The research team included a balance of investigators with expertise relevant to the clinical content area of the report (perinatal and postpartum psychiatry, general obstetrics and gynecology, maternal/fetal medicine, general pediatrics) and/or methodological expertise in epidemiology, screening, decision modeling, and the conduct of systematic reviews. 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. Articles meeting our eligibility criteria were included for data abstraction. Relevant review articles and meta-analyses 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 article and the associated completed abstraction form 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 gave 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, because screening, diagnostic, and 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 and the Medical Test Guide. 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 lists 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 and the Medical Test Guide. In this context, “bias” refers to the degree to which a study’s results are due to aspects of the study design (choice of population, allocation of treatment, uneven distribution of risk factors, etc.) rather than the specific factor (risk factor or exposure, screening test, treatment, etc.) of interest. 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 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.
Table 4. Definitions of overall quality ratings

<table>
<thead>
<tr>
<th>Quality Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>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.</td>
</tr>
<tr>
<td>Fair</td>
<td>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.</td>
</tr>
<tr>
<td>Poor</td>
<td>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.</td>
</tr>
</tbody>
</table>

For studies assessing screening test performance (KQs 1, 2, and 3), we used QUADAS-2 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 then 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. In such instances if a meta-analysis was appropriate, we planned to use random-effects models to synthesize the available evidence quantitatively using Comprehensive Meta-Analysis software (Version 2; Biostat, Englewood, NJ). This approach includes testing for heterogeneity using graphical displays and test statistics (Q and \( I^2 \) statistics, recognizing that the ability of statistical methods to detect heterogeneity may be limited). For comparison purposes, we also planned to perform fixed-effect meta-analyses. Our presentation of meta-analysis results typically includes
summary estimates, standard errors, and exact confidence intervals (CIs). Unfortunately, the available evidence did not support meta-analysis of patient-centered or adverse event outcomes.

Test performance was summarized using sensitivity and specificity. 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.

If test performance studies were conceptually homogeneous, we planned to use random-effects bivariate meta-analysis using SAS version 9.2 (SAS Institute Inc., Cary, NC) to compute summary estimates of performance. 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. We intended to evaluate statistical heterogeneity by inspecting forest plots and computing $Q$ and $I^2$ statistics.

Since the Q test is underpowered, we planned to 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.

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 to perform 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. As with the patient-centered and adverse event outcomes, the available evidence did not support meta-analyses of test performance or subgroup data.

We also adapted an existing simulation model of pregnancy and neonatal outcomes 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, 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 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.\(^8\) 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\(^32,81\) and Medical Test Guide.\(^33\) 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

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

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\textsuperscript{32,81} and Medical Test Guide;\textsuperscript{33} 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., 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 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 (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 were invited to provide external peer review of this draft report; AHRQ and an associate editor also provided comments. The draft report was posted on AHRQ’s Web site for public comment for 4 weeks, from July 31 to August 28, 2012. We have addressed reviewer comments, revising the report as appropriate, and have documented our responses in a disposition of comments report available on the AHRQ Web site. A list of peer reviewers is given in the preface of this 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 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 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.
Some studies were relevant to more than one KQ.

5,059 citations identified by literature search:
- PubMed: 2,525
- Cochrane: 7
- Embase: 1,093
- PsycINFO: 1,434

1,528 duplicates

Manual searching: 154

3,685 citations identified

2,392 abstracts excluded

1,293 passed abstract screening

1,248 articles excluded:
- Full-text unavailable: 1
- Not a full publication (abstract only), or not original peer-reviewed data: 206
- Not a randomized trial, or not an observational study of appropriate design with sample size ≥100: 72
- Not a population of interest: 167
- Study locations did not include at least one high-income economy: 33
- No interventions of interest, or timing of intervention not within range: 129
- No comparator of interest: 129
- No outcomes of interest: 476
- Published prior to January 1, 2004: 35

45 articles representing 40 studies passed full-text screening

Data abstracted for 40 studies:
- KQ 1: 20 articles (18 studies)
- KQ 2: 16 articles (15 studies)
- KQ 3: 2 articles (2 studies)
- KQ 4: 5 articles (5 studies)
- KQ 5: 1 article (1 study)
- KQ 6: 6 articles (6 studies)

*aSome studies were relevant to more than one KQ.*
Description of Included Studies

Overall, we included 40 studies represented by 45 publications: 18 studies were relevant to KQ 1, 15 studies to KQ 2, 2 studies to KQ 3, 5 studies to KQ 4, 1 study to KQ 5, and 6 studies to KQ 6. Studies were conducted in Europe (33%), the United States or Canada (33%), the UK (15%), Asia (10%), Australia or New Zealand (7%), and other locations (2%). Further details on the studies included for each KQ are provided in the relevant results sections, below, and in Appendix E. Forty of these studies reported results for women, while one reported on the test characteristics of the Edinburgh Postnatal Depression Scale (EPDS) in fathers.

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 117 trial records. A single reviewer identified six of these records as potentially relevant. Of note, no other potentially relevant records beyond these six were identified from a parallel search of the WHO ICTRP registry platform. Three of the six identified studies of interest are complete; 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 to 90% range for most screening tests across studies.
• In two studies at low risk of bias, a two-question screen had a sensitivity of 100 percent (95% confidence intervals [CIs], 93.3 to 100% and 83.3 to 100%), with specificities of 44.0 percent (95% CI, 39.5 to 48.8%) and 64.5 percent (53.7 to 75.2%), 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. In another study comparing the EPDS to the Mood Spectrum Self-Report (MOODS-SR), suicidal ideation was more common in the EPDS than the MOODS-SR, although formal tests of agreement were not performed.

• In one large study at high risk of bias, performance characteristics for self-administered instruments designed for screening (the EPDS and Beck Depression Inventory [BDI] were similar to those of an interview-based instrument typically used in diagnostic settings (the Hamilton Rating Scale for Depression [HRSD]).

• 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 for major depression are generally in the 80–90 percent range at commonly used thresholds; as sensitivity increases with choice of thresholds, specificity decreases and vice versa. 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.

• One study judged at high risk of bias reported test characteristics for the EPDS in fathers tested 7 weeks after birth (sensitivity 89.5%; 95% CI, 66.9 to 98.7%; specificity 78.2%; 95% CI, 71.3 to 84.2% at a threshold of 10) that are similar to those reported for mothers.

**Description of Included Studies**

We identified 18 studies that met the inclusion criteria for KQ 1. 1. Two of these studies are each represented by two included publications. The 2004 2 and 2006 3 publications by Felice et al. describe results for the same study population, as do the 2009 4 and 2011 5 publications by Gjerdingen et al. All 18 studies 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, 6,7,8,9 and one each in Australia, New Zealand, Asia, and Canada. Ten studies were judged to have a high risk of biased results; the remainder were judged to be at low risk. One of the 18 studies focused on fathers and the rest on mothers.

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 two studies were relevant to KQ 1b.
Eleven studies provided sensitivity and specificity 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 various versions of the Patient Health Questionnaire (PHQ), and one each on the Antenatal Risk Questionnaire, the 17- and 21-Item Hamilton Rating Scale for Depression (HRSD-17 and HRSD-21), and the Leverton Questionnaire.

**Detailed Synthesis**

**Edinburgh Postnatal Depression Scale (EPDS)**

Eleven studies provided sensitivity and specificity data for major depression for the EPDS.\(^8\) Studies varied in the threshold used to define a positive screening test. There was a clear trend toward 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 that used thresholds of 8\(^9\) and 13\(^8\). Even though, as expected, confidence intervals were considerably narrower for specificity, there was again considerable overlap across thresholds from 10 through 12.

Of note, one of these studies\(^8\) was performed in 189 fathers, with test characteristics (sensitivity 89.5%; 95% CI, 66.9 to 98.7%; specificity 78.2%; 95% CI, 71.3 to 84.2%) quite similar to those observed in women at the same threshold as seen in Figures 3 and 4. The study was judged to be at high risk of bias because of relatively low participation, and the reference standard was preferentially applied to men with high scores on the screening test, which creates potential for ascertainment bias (overestimation of sensitivity).

**Figure 3. Sensitivity of the EPDS at various thresholds**

CI = confidence interval; EPDS = Edinburgh Postnatal Depression Scale

Note: Data from Ekeroma 2012\(^10\) are for the Samoan subgroup.
Figure 4. Specificity of the EPDS at various thresholds

CI = confidence interval; EPDS = Edinburgh Postnatal Depression Scale

Note: Data from Ekeroma 2012 are for the Samoan subgroup.

Postpartum Depression Screening Scale (PDSS)

Four studies provided sensitivity and specificity data for the PDSS across a range of thresholds.84-86,98 Of note, Beck et al.84 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 study84 (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 were wider for sensitivity than for specificity, and there was a clear trend toward 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

CI = confidence interval; PDSS = Postpartum Depression Screening Scale

Figure 6. Specificity of the PDSS at various thresholds

CI = confidence interval; PDSS = Postpartum Depression Screening Scale
Other Tests

There were only one to two studies providing data for each of the other screening tests of interest in this review. Table 6 summarizes sensitivity/specificity results for these studies.

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

<table>
<thead>
<tr>
<th>Screening Test</th>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antenatal Risk Questionnaire</td>
<td>Austin, 2011⁸²</td>
<td>78.2% (65.0 to 88.2%)</td>
<td>47.1% (40.3 to 53.9%)</td>
</tr>
<tr>
<td>BDI</td>
<td>Ji, 2011⁹⁶</td>
<td>82.8% (73.9 to 90.0%)</td>
<td>82.1% (79.2 to 85.0%)</td>
</tr>
<tr>
<td>BDI (BDI-1A)</td>
<td>Csatorndai, 2009⁸⁷</td>
<td>91.1% (86.1 to 95.7%)</td>
<td>91.5% (88.7 to 93.9%)</td>
</tr>
<tr>
<td>BDI-II</td>
<td>Pereira, 2010⁹⁸</td>
<td>88.2% (73.3 to 100%)</td>
<td>87.7% (84.6 to 90.8%)</td>
</tr>
<tr>
<td>BDI-II</td>
<td>Chaudron, 2010⁸⁵</td>
<td>74.0% (63.9 to 84.1%)</td>
<td>79.7% (72.6 to 86.8%)</td>
</tr>
<tr>
<td>HRSD-17</td>
<td>Ji, 2011⁹⁶</td>
<td>81.0% (65.9 to 91.4%)</td>
<td>80.9% (76.4 to 85.0%)</td>
</tr>
<tr>
<td>HRSD-21</td>
<td>Ji, 2011⁹⁶</td>
<td>81.0% (65.9 to 91.4%)</td>
<td>75.7% (70.8 to 80.1%)</td>
</tr>
<tr>
<td>Leverton Questionnaire</td>
<td>Csatorndai, 2009⁸⁷</td>
<td>95.2% (90.4 to 98.1%)</td>
<td>91.3% (88.4 to 93.7%)</td>
</tr>
<tr>
<td>PHQ-2 yes/no</td>
<td>Gjerdingen, 2009⁹¹</td>
<td>100% (93.3 to 100%)</td>
<td>44.0% (39.5 to 48.8%)</td>
</tr>
<tr>
<td>PHQ-2 yes/no</td>
<td>Mann, 2012¹⁰⁰</td>
<td>100% (83.3 to 100%)</td>
<td>64.5% (53.7 to 75.2%)</td>
</tr>
<tr>
<td>PHQ-2 Likert</td>
<td>Gjerdingen, 2009⁹¹</td>
<td>84.4% (70.5 to 93.5%)</td>
<td>78.7% (74.7 to 82.4%)</td>
</tr>
<tr>
<td>PHQ-9-simple</td>
<td>Gjerdingen, 2009⁹¹</td>
<td>82.2% (68.0 to 92.0%)</td>
<td>83.8% (80.1 to 87.0%)</td>
</tr>
<tr>
<td>PHQ-9 complex</td>
<td>Gjerdingen, 2009⁹¹</td>
<td>66.7% (52.9 to 80.4%)</td>
<td>91.5% (89.0 to 94.1%)</td>
</tr>
</tbody>
</table>

BDI = Beck Depression Inventory; BDI-II = Beck Depression Inventory-II; 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. The BDI and the 25-item Leverton Questionnaire had both sensitivity and specificity above 90 percent in a Hungarian validation study.⁸⁷ This study was rated as having a low risk of bias, and included 1,552 subjects.

Notably, 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.⁹¹,⁹²,¹⁰⁰

Within-Study Comparisons

Three studies compared different instruments in the same population (Figures 7 and 8).⁸⁵,⁸⁶,⁹⁶ The largest⁹⁶ compared the EPDS, BDI, HRSD-17, and HRSD-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 HRSD is an interview-based instrument and is generally not considered a screening test—providing some evidence that self-administered instruments offer comparable performance to interview-based instruments. The two studies directly comparing the EPDS and PDSS 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**

*BDI = Beck Depression Inventory; BDI-II = Beck Depression Inventory-II; 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 two studies, one performed in Scotland\textsuperscript{94} and one in Italy,\textsuperscript{99} that evaluated the suicidal ideation component of the EPDS.

Howard et al.\textsuperscript{94} performed a nested cohort study within an RCT for postpartum depression. The prevalence of suicidal ideation (Item 10) on the EPDS at 6 to 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).

Mauri et al.,\textsuperscript{99} as part of a larger cohort study of perinatal depression, compared the prevalence of suicidal ideation on the EPDS to that on the Mood Spectrum Self-Report (MOODS-SR) at 1, 3, 6, 9, and 12 months postpartum. Period prevalence for suicidal ideation during the postpartum period was 8.6 percent (95% CI, 7.4 to 9.8%) for the EPDS and 4.3 percent (95% CI, 3.4 to 5.2%) for the MOODS-SR. Point prevalence generally declined over time, with estimates based on the EPDS consistently higher than for the MOODS-SR at every time point (prevalence at 1 month postpartum for EPDS was 2.7%, 95% CI, 2.1 to 3.3%; for MOODS-SR it was 1.2%, 95% CI, 0.8 to 1.6%). Formal statistics on agreement between the two instruments were not provided, and insufficient data were provided to allow calculation of kappa.
statistics. Twenty-five percent of those with suicidal ideation on the EPDS also would have also met criteria on the MOODS-SR. In multivariate analysis, suicidal ideation on the EPDS was associated with major depression during pregnancy or the postpartum period.

**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.
- We did not identify any studies that explicitly and directly compared the predictive values of screening instruments in different populations; only one study reported on potential differences in test sensitivity and specificity based on the presence of a specific characteristic or risk factor.
- 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.
- All of the associations noted above are consistent with findings of studies published prior to our search dates and noted in recent reviews.
- 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 16 articles describing 15 unique studies that met the inclusion criteria for KQ 2.95,96,102-115 (The 2005109 and 2008108 publications by Chee et al. described results for the same study population.) Three were from the United States,96,102,105 seven were from Europe,95,103,110-114 two (three publications) were Asian,108,109,115 and there was one study each from the UK,104 Australia,107 and Israel.106 Applying QUADAS-2 criteria across the studies applicable to KQ 2, 2 studies were judged to be low risk of bias,104,115 10 high risk of bias,95,96,102,105,106,110-112,114 and 3 unclear risk of bias.103,107,113 We did not identify any studies relevant to KQ 2b. One study judged to be at high risk of bias96 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; in the list of variables, the bold text refers to the specific predictor for which measures of association were presented (e.g., maternal age). 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 study95 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

<table>
<thead>
<tr>
<th>Study Total N</th>
<th>Quality</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akincigil, 2010102</td>
<td>4,348</td>
<td>Maternal age ≤21 years</td>
<td>Referent</td>
<td>–</td>
<td>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</td>
</tr>
<tr>
<td>Maternal age 22–24</td>
<td>1.08</td>
<td>NS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal age ≥25 years</td>
<td>0.81</td>
<td>NS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boyce, 2005107</td>
<td>425</td>
<td>Maternal age</td>
<td>0.96</td>
<td>0.88 to 1.05</td>
<td>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</td>
</tr>
</tbody>
</table>
Table 7. Maternal demographic risk factors for postpartum major depression (continued)

<table>
<thead>
<tr>
<th>Study Total N</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siu, 2012115</td>
<td>Maternal age</td>
<td>28.8 years vs. 30.3 years</td>
<td>P=0.001 in univariate analysis, NS when adjusted for other variables</td>
<td>t-test only; age not included in final backward stepwise logistic regression; final model marital dissatisfaction, poor relationship with mother-in-law, antenatal depressive symptomatology, and anxiety-prone-personality</td>
</tr>
<tr>
<td>569</td>
<td>Low risk of bias</td>
<td>Maternal age, continuous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Akincigil, 2010102</td>
<td>Education: Less than HS</td>
<td>Referent</td>
<td>–</td>
<td>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</td>
</tr>
<tr>
<td>4,348</td>
<td>Education: HS or equivalent</td>
<td>1.02</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>High risk of bias</td>
<td>Education: Some college</td>
<td>1.23</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>Education: College or more</td>
<td>1.15</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boyce, 2005107</td>
<td>Education (not specified)</td>
<td>1.04</td>
<td>0.84 to 1.27</td>
<td>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</td>
</tr>
<tr>
<td>425</td>
<td>Unclear risk of bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Akincigil, 2010102</td>
<td>Income/poverty ratio: &lt;100%</td>
<td>Referent</td>
<td>–</td>
<td>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</td>
</tr>
<tr>
<td>4,348</td>
<td>Income/poverty ratio: 100–300%</td>
<td>1.17</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>High risk of bias</td>
<td>Income/poverty ratio: &gt;300%</td>
<td>0.82</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>Andersson, 2006103</td>
<td>Type of employment: Laborer</td>
<td>Referent</td>
<td>–</td>
<td>Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, first-trimester BMI</td>
</tr>
<tr>
<td>650</td>
<td>Unclear risk of bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jardri, 200695</td>
<td>Unemployed</td>
<td>2.8</td>
<td>1.1 to 4.9</td>
<td>History of postpartum depression, history of depression, preterm birth, stopping breastfeeding, multiple gestation, postpartum complications, employment status</td>
</tr>
<tr>
<td>363</td>
<td>High risk of bias</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

*Total N analyzed for study; may not reflect N assessed for indicated risk factor.

Table 8 summarizes the available data on potential risk factors relating to obstetric history. As shown there, having a preterm or very low birthweight baby were both significantly...
associated with postpartum depression. In another study not shown in Table 8, having a termination of pregnancy for a severe fetal malformation or chromosomal abnormality in the second or third 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, with a positive but nonsignificant association in another. One study reported a significant increase in risk of depression with prenatal smoking, while another reported a nonsignificant decrease in risk among smokers.

Table 8. Obstetric history risk factors for postpartum depression

<table>
<thead>
<tr>
<th>Study Total N</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akincigil, 2010</td>
<td>Very low birthweight infant</td>
<td>1.63</td>
<td>P&lt;0.1</td>
<td>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</td>
</tr>
<tr>
<td>Jardri, 2006</td>
<td>Delivery prior to 37 weeks</td>
<td>4.5</td>
<td>1.4 to 14.6</td>
<td>History of postpartum depression, history of depression, preterm birth, stopping breastfeeding, multiple gestation, postpartum complications, employment status</td>
</tr>
<tr>
<td>Akincigil, 2010</td>
<td>Parity: ≥3 children</td>
<td>Referent</td>
<td>–</td>
<td>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</td>
</tr>
<tr>
<td>Andersson, 2006</td>
<td>Parity: Nulliparous</td>
<td>Referent</td>
<td>–</td>
<td>Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, first-trimester BMI</td>
</tr>
<tr>
<td>Akincigil, 2010</td>
<td>Prenatal tobacco use</td>
<td>1.23</td>
<td>P&lt;0.1</td>
<td>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</td>
</tr>
</tbody>
</table>
### Table 8. Obstetric history risk factors for postpartum depression (continued)

<table>
<thead>
<tr>
<th>Study</th>
<th>Total N</th>
<th>Quality</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andersson, 2006103</td>
<td>650</td>
<td>Unclear risk of bias</td>
<td>Smoker</td>
<td>0.28</td>
<td>0.03 to 2.45</td>
<td>Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, first-trimester BMI</td>
</tr>
<tr>
<td>Akincigil, 2010102</td>
<td>4,348</td>
<td>High risk of bias</td>
<td>Prenatal alcohol use</td>
<td>1.14</td>
<td>NS</td>
<td>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</td>
</tr>
</tbody>
</table>

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

*Total N analyzed for study; may not reflect N assessed for indicated risk factor.

One study96 estimated sensitivity and specificity for the BDI, EPDS, HRSD-17, and HRSD-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

<table>
<thead>
<tr>
<th>Timing, Instrument, and Gravidity</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point Estimate</td>
<td>Lower 95% CI</td>
</tr>
<tr>
<td>Discharge to 8 Weeks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BDI</td>
<td>Primigravid</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>88.5%</td>
</tr>
<tr>
<td>EPDS</td>
<td>Primigravid</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>81.0%</td>
</tr>
<tr>
<td>HRSD-17</td>
<td>Primigravid</td>
<td>75.0%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>82.4%</td>
</tr>
<tr>
<td>HRSD-21</td>
<td>Primigravid</td>
<td>75.0%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>85.3%</td>
</tr>
<tr>
<td>8 Weeks to 6 Months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BDI</td>
<td>Primigravid</td>
<td>86.7%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>80.7%</td>
</tr>
<tr>
<td>EPDS</td>
<td>Primigravid</td>
<td>87.5%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>76.2%</td>
</tr>
<tr>
<td>HRSD-17</td>
<td>Primigravid</td>
<td>90.5%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>77.2%</td>
</tr>
<tr>
<td>HRSD-21</td>
<td>Primigravid</td>
<td>85.7%</td>
</tr>
<tr>
<td></td>
<td>Multigravid</td>
<td>80.4%</td>
</tr>
</tbody>
</table>

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\textsuperscript{102} and a history of chronic illness outside of pregnancy\textsuperscript{103} 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;\textsuperscript{112} 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

<table>
<thead>
<tr>
<th>Study</th>
<th>Total N\textsuperscript{a}</th>
<th>Quality</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akincigil, 2010\textsuperscript{102}</td>
<td>4,348</td>
<td>High risk of bias</td>
<td>Self-rated health status: Great/very good</td>
<td>Referent</td>
<td>–</td>
<td>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</td>
</tr>
<tr>
<td>Andersson, 2006\textsuperscript{103}</td>
<td>650</td>
<td>Unclear risk of bias</td>
<td>History of chronic illness</td>
<td>2.71</td>
<td>2.36 to 19.14</td>
<td>Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, first-trimester BMI</td>
</tr>
<tr>
<td>Andersson, 2006\textsuperscript{103}</td>
<td>650</td>
<td>Unclear risk of bias</td>
<td>1\textsuperscript{st} trimester BMI: 18.5–24.9</td>
<td>Referent</td>
<td>–</td>
<td>Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, first-trimester BMI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1\textsuperscript{st} trimester BMI: 25–29.9</td>
<td>2.17</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1\textsuperscript{st} trimester BMI: ≥30</td>
<td>No cases</td>
<td>NR</td>
<td></td>
</tr>
</tbody>
</table>

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

\textsuperscript{a}Total N analyzed for study; may not reflect N assessed for indicated risk factor.

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.\textsuperscript{107,113}
Table 11. Psychiatric history risk factors for postpartum major depression

<table>
<thead>
<tr>
<th>Study</th>
<th>Total N</th>
<th>Quality</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>History of Depression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bloch, 2005⁹⁶</td>
<td>244</td>
<td>High risk of bias</td>
<td>History of premenstrual dysphoric disorder</td>
<td>NR</td>
<td>0.048</td>
<td>Postpartum mood, <strong>history of depression</strong>, history of mood symptoms while using oral contraceptives, EPDS score, <strong>history of premenstrual dysphoric disorder</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>History of major depressive disorder</td>
<td>NR</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Chee, 2005⁹⁷</td>
<td>278</td>
<td>High risk of bias</td>
<td>History of depression</td>
<td>4.91</td>
<td>1.08 to 22.3</td>
<td>Education, marital dissatisfaction, confinement, low instrumental support⁸ history of depression</td>
</tr>
<tr>
<td>Garcia-Esteve, 2008¹⁰⁰</td>
<td>334</td>
<td>High risk of bias</td>
<td>History of depression</td>
<td>3.67</td>
<td>1.63 to 8.27</td>
<td>Family caregiver role, poor partner relationship, low social support during pregnancy</td>
</tr>
<tr>
<td>Jardri, 2006⁹⁵</td>
<td>363</td>
<td>High risk of bias</td>
<td>History of postpartum depression</td>
<td>4.3</td>
<td>1.7 to 10.9</td>
<td>History of postpartum depression, <strong>history of depression</strong>, preterm birth, stopping breastfeeding, multiple gestation, postpartum complications</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>History of depression</td>
<td>4.4</td>
<td>2.2 to 9.0</td>
<td></td>
</tr>
<tr>
<td>Verkerk, 2005¹¹³</td>
<td>277</td>
<td>Unclear risk of bias</td>
<td>History of depression</td>
<td>3.08</td>
<td>1.10 to 8.63</td>
<td>Covariates not reported; results presented here for 3 months postpartum, similar findings for 6 and 12 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Family history of depression</td>
<td>1.60</td>
<td>0.67 to 3.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Depression during pregnancy</td>
<td>2.10</td>
<td>0.82 to 5.37</td>
<td></td>
</tr>
<tr>
<td>Siu, 2012¹⁰⁵</td>
<td>569</td>
<td>Low risk of bias</td>
<td>History of depression</td>
<td>3.59</td>
<td>2.27 to 5.68</td>
<td>Marital dissatisfaction, poor relationship with mother-in-law, <strong>antenatal depressive symptomatology</strong>, and anxiety-prone-personality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Depression during pregnancy</td>
<td>3.9</td>
<td>3.04 to 4.99</td>
<td></td>
</tr>
<tr>
<td><strong>Other Psychiatric Disorders</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andersson, 2006¹⁰³</td>
<td>650</td>
<td>Unclear risk of bias</td>
<td>History of psychiatric disorder</td>
<td>6.72</td>
<td>2.36 to 19.14</td>
<td>Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, first-trimester BMI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boyce, 2005¹⁰⁷</td>
<td>425</td>
<td>Unclear risk of bias</td>
<td>Past psychiatric history</td>
<td>2.74</td>
<td>0.60 to 12.45</td>
<td>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</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vulnerable personality</td>
<td>2.82</td>
<td>1.06 to 7.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low organized/responsive personality</td>
<td>3.69</td>
<td>1.26 to 10.8</td>
<td></td>
</tr>
</tbody>
</table>
Table 11. Psychiatric history risk factors for postpartum major depression (continued)

<table>
<thead>
<tr>
<th>Study</th>
<th>Total N</th>
<th>Quality</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mauri, 2010</td>
<td>500</td>
<td>High risk of bias</td>
<td>Any anxiety disorder</td>
<td>2.4</td>
<td>1.1 to 5.7</td>
<td>Other covariates not specified</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Panic disorder</td>
<td>8.0</td>
<td>2.75 to 23.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Social phobia</td>
<td>7.65</td>
<td>2.64 to 22.2</td>
<td></td>
</tr>
<tr>
<td>Verkerk, 2005</td>
<td>277</td>
<td>Unclear risk of bias</td>
<td>High neuroticism/high introversion</td>
<td>3.08</td>
<td>1.10 to 8.63</td>
<td>Covariates not reported; results presented here for 3 months postpartum, similar findings for 6 and 12 months</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High neuroticism/low introversion</td>
<td>1.58</td>
<td>0.51 to 4.93</td>
<td></td>
</tr>
<tr>
<td>Siu, 2012</td>
<td>569</td>
<td>Low risk of bias</td>
<td>Antenatal stressful life events</td>
<td>2.56</td>
<td>1.84 to 3.57</td>
<td>Marital dissatisfaction, poor relationship with mother-in-law, antenatal depressive symptomatology, and anxiety-prone personality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Anxiety-prone personality</td>
<td>2.14</td>
<td>1.79 to 2.56</td>
<td></td>
</tr>
</tbody>
</table>

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

Total N analyzed for study; may not reflect N assessed for indicated risk factor.

Instrumental support is the provision of financial assistance, material goods, or services, also called “tangible support.”

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Total N</th>
<th>Quality</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akincigil, 2010</td>
<td>4,348</td>
<td>High risk of bias</td>
<td>Married</td>
<td>Referent</td>
<td>--</td>
<td>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</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cohabiting</td>
<td>0.98</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Visiting</td>
<td>0.97</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>Andersson, 2006</td>
<td>650</td>
<td>Unclear risk of bias</td>
<td>No relation</td>
<td>0.64</td>
<td>P&lt;0.1</td>
<td>Age, socioeconomic status, smoking status, snuff use, parity, alcohol use, chronic disease, history of psychiatric disorder, first-trimester BMI, marital status</td>
</tr>
<tr>
<td>Siu, 2012</td>
<td>569</td>
<td>Low risk of bias</td>
<td>Single (compared with married)</td>
<td>26.4</td>
<td>4.14 to 168.3</td>
<td>Marital dissatisfaction, poor relationship with mother-in-law, antenatal depressive symptomatology, and anxiety-prone personality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unmarried</td>
<td>2.21</td>
<td>1.28 to 3.83</td>
<td></td>
</tr>
</tbody>
</table>

Marital Status

39
Table 12. Relationship and social support risk factors for postpartum major depression (continued)

<table>
<thead>
<tr>
<th>Study Total N</th>
<th>Risk Factor</th>
<th>RR/OR</th>
<th>95% CI or P Value</th>
<th>Variables Included in Multivariate Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Other Factors</td>
<td></td>
<td></td>
<td>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</td>
</tr>
<tr>
<td>Akincigil, 2010</td>
<td>History of violence/abuse</td>
<td>1.36</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>4,348</td>
<td>Better relationship quality</td>
<td>0.89</td>
<td>p&lt;0.05</td>
<td>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</td>
</tr>
<tr>
<td>High risk of bias</td>
<td>Disagreement about pregnancy</td>
<td>1.41</td>
<td>p&lt;0.05</td>
<td>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</td>
</tr>
<tr>
<td>Barnes, 2009</td>
<td>More social support</td>
<td>0.89</td>
<td>0.80 to 0.98</td>
<td>Education, occupation, age, ethnicity, marital status, number of children, stress, social support, depression at 2 months</td>
</tr>
<tr>
<td>250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclear risk of bias</td>
<td>Dissatisfaction with partner</td>
<td>1.38</td>
<td>0.23 to 8.19</td>
<td>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</td>
</tr>
<tr>
<td>Boyce, 2005</td>
<td>Worsening relationship</td>
<td>2.45</td>
<td>0.78 to 6.47</td>
<td>Education, marital dissatisfaction, confinement, low instrumental support, history of depression</td>
</tr>
<tr>
<td>425</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chee, 2005</td>
<td>Marital dissatisfaction</td>
<td>9.42</td>
<td>2.19 to 40.52</td>
<td>Education, marital dissatisfaction, confinement, low instrumental support, history of depression</td>
</tr>
<tr>
<td>278</td>
<td>Negative “confinement” experience (restricted activities per cultural norms in different ethnic communities)</td>
<td>19.41</td>
<td>2.03 to 185.5</td>
<td>Education, marital dissatisfaction, confinement, low instrumental support, history of depression</td>
</tr>
<tr>
<td>High risk of bias</td>
<td>Low instrumental support</td>
<td>23.43</td>
<td>3.68 to 149.16</td>
<td>Education, marital dissatisfaction, confinement, low instrumental support, history of depression</td>
</tr>
<tr>
<td>Chee, 2008</td>
<td>Less emotional support</td>
<td>1.92</td>
<td>1.12 to 3.68</td>
<td>Education, marital dissatisfaction, confinement, low instrumental support, history of depression</td>
</tr>
<tr>
<td>307</td>
<td>3 or more nonscheduled pediatric visits</td>
<td>2.87</td>
<td>1.41 to 5.85</td>
<td></td>
</tr>
<tr>
<td>Garcia-Esteve, 2008</td>
<td>Family caregiver role</td>
<td>4.39</td>
<td>1.10 to 17.4</td>
<td>Family caregiver role, poor partner relationship, low social support during pregnancy</td>
</tr>
<tr>
<td>Poor partner relationship</td>
<td>4.24</td>
<td>1.38 to 13.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low social support</td>
<td>4.06</td>
<td>1.47 to 11.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siu, 2012</td>
<td>Poor marital relationship</td>
<td>8.27</td>
<td>5.06 to 13.5</td>
<td>Marital dissatisfaction, poor relationship with mother-in-law, antenatal depressive symptomatology, and anxiety-prone personality</td>
</tr>
<tr>
<td>115</td>
<td>Poor relationship with mother-in-law</td>
<td>3.93</td>
<td>3.05 to 5.04</td>
<td>Marital dissatisfaction, poor relationship with mother-in-law, antenatal depressive symptomatology, and anxiety-prone personality</td>
</tr>
<tr>
<td>Felt stress in childcare</td>
<td>2.20</td>
<td>1.88 to 2.57</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

*aTotal N analyzed for study; may not reflect N assessed for indicated risk factor.

*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.\textsuperscript{96,116} No studies were identified that met the inclusion criteria for KQ 3b or KQ 3c.

The first study, a prospective investigation of maternal mental illness conducted at a single academic center in the United States,\textsuperscript{96} enrolled women prior to 28 weeks gestation and followed them 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\textsuperscript{116} was a single-center prospective investigation conducted in Dublin, Ireland. This study 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.,\textsuperscript{96} 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

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

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.,116 625 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 for this study 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 to 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 with 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 U.K.–based RCT found that there is benefit to the overall mental health in mothers in screening, with initial treatment in screen positives, by health visitors at 6 weeks postnatally.
- One fair-quality U.S.–based RCT found that a practice-level intervention to improve screening between 5 and 12 weeks postpartum compared with usual care in family medicine clinics led to lower levels of depressive symptoms at 6 and 12 months. A smaller, fair-quality Hong Kong–based RCT also found improvement in depressive symptoms, but no difference in overall mental health scores.
- None of the three RCTs which included a measure of parental stress (the Parental Stress Inventory [PSI]) showed improvement in this measure with screening and treatment of depressive symptoms. The included studies do not allow an assessment of the comparative benefits of screening for postpartum depression by screening instrument, timing, or setting.
- None of the studies included outcomes for fathers.

Description of Included Studies

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. Among the four RCTs, one was rated as poor quality, two as fair, and one as 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 sizes ranged from 99 recruited at a single site to 4,084 enrolled from 101 practices. Two studies were conducted in the United States. The remaining studies were conducted in the UK, Europe, and Asia.

Detailed Synthesis

A good-quality cluster-randomized trial evaluated screening in 101 general practices in Trent, England. 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 postpartum 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 health visitors in the intervention practices, those in the control practices 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. Parenting stress was also measured at 6, 12, and 18 months using the PSI-Short Form (PSI-SF).
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 postpartum 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 in the trial (n=2,659), 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.

Among those who scored 12 or greater on the EPDS at 6 weeks, the unadjusted score on the PSI-SF was statistically significantly higher (p=0.001) in the intervention (n=211) compared with the control group (n=106), with a difference of 9.2 (95% CI, 4.8 to 13.7). At 12 months, the difference between the intervention (n=156) and the control groups (n=90) was 8.0 (95% CI, 3.1 to 13.0), and at 18 months, the difference between the intervention (n=82) and the control groups (n=46) was 9.1 (95% CI, 1.1 to 17.4).

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 percent had followup at 6 months; in the intervention group, 82.8 percent had followup at 6 weeks and 68.4 percent 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.

The earliest of the five studies identified 201 women receiving public assistance and at high risk for postpartum depression based on a 17-item depression risk survey. 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 at 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 quasi-experimental posttest study 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.

In a fair-quality RCT, 28 of 33 primary care practices completed a study in which they were randomized into two arms: usual care with training about postpartum depression and an active arm with more extensive training and implementation of the EPDS for screening of women between 5 and 12 weeks postpartum. Initial followup within the practices for those with an elevated EPDS was with a practice-administered PHQ-9. Women in the usual care practices completed the EPDS and PHQ-9, which were submitted to a central study site instead of to the clinicians in the practice. Overall, there were 990 women in the usual care group and 1,353 in the intervention group. Among these, 255 (26%) in the usual care group and 399 (29%) in the intervention group had an EPDS ≥10 or PHQ-9 ≥10. Overall, women in the intervention group with elevated EPDS scores were more likely to be diagnosed with depression than those in the usual care group (66% vs. 41%; p=0.001). Similarly, those in the intervention group with elevated EPDS scores were more likely to receive medication (56% vs. 35%; p<0.001), counseling (20% vs. 11%; p=0.02), or both (60% vs. 37%; p<0.0001). At 12 months, the adjusted odds ratio for a 5-point or greater decrease in the PHQ-9 in the intervention group compared with the control group was 1.82 versus 1.74 (p<0.001). Interestingly, in the same multivariate analysis, higher scores on the Parenting Stress Index (PSI) at baseline were associated with significantly lower chance of a 5-point or greater decrease in the PHQ-9 in both the intervention and control groups. There was no significant difference in changes in scores at 12 months for either the PSI or the Dyad Adult Satisfaction (DAS-6) scale (a measure of relationship satisfaction). This study was considered to be fair quality primarily because the main outcome of depression was based on self-reported symptoms and a chart audit rather than a standardized clinical assessment.

Another fair-quality RCT evaluated the effectiveness of screening for postpartum depression with the EPDS compared with no screening. 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 (n=231), women attending the centers were screened using 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. 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 (76.1%) received treatment. In the control group, 14 women were assessed as having postpartum depression, and 10 (71.4%) received treatment. Based on an intention-to-treat 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). This difference in EPDS scores 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). 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]). However, children of screened women had more visits to the doctors than did children of women in the usual care group (p=0.039), even after adjusting for possible differences in baseline health status. 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 five 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 suggests that screening provides a benefit to overall mental health in mothers based on the SF-12 MCS. Similarly, two fair-quality RCTs found a benefit from screening with the EPDS, with reduced levels of depressive symptoms. Although there is no direct evidence of differences in setting, it is notable that the two studies with the greatest effect sizes tested strategies where treatment was provided within the same setting as screening (home visitation or family practice clinic), rather than a setting where further management of women with positive screening results required referral to a different provider. It is unclear how these observed benefits translate into improved quality of life, family functioning, or other health outcomes, especially since the three studies that collected data on a measure of parental functioning found no difference between groups. 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 with 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, even after adjustment for baseline health, than did women in a control group. It is unclear whether this difference represents overutilization on the part of the screened group, or underutilization by the unscreened group.

**Description of Included Study**

Only one study met the inclusion criteria for KQ 5; this was a fair-quality RCT conducted in Asia. 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, children of screened women had more doctor visits (mean 2.39; 95% CI, 2.07 to 2.7) compared with 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 statistically 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. Although adjustment of baseline health status suggests that these differences in visit utilization may reflect differences in the appropriateness of the visits, there is no evidence to suggest whether any differences were due to overutilization among the screened or underutilization among the unscreened.

**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 for women who screened positive for postpartum depression were low, ranging from 0–30 percent, except for one trial where screening, diagnosis, and treatment were all conducted within a primary care setting, and where 60 percent rates of treatment were achieved.
- Evidence on the effect of timing of screening on referral rates was mixed:
  - One good-quality cross-sectional study found that women who screened positive for 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.
  - Conversely, a poor-quality prospective cohort study conducted in a private-practice setting found that, while all women who screened positive for depression received a referral for followup care, no women who screened positive during pregnancy sought care, and only 18 percent of those who screened positive at 6 weeks postpartum sought care.
- A fair quality RCT demonstrated high levels of receipt of appropriate services among primary care practices where screening and treatment occurred within the same setting. These levels were substantially higher than those reported in other settings and were associated with significant improvement in depressive symptoms, but the study design precludes drawing inferences about the comparative effectiveness of screening in this type of setting compared with other settings.

**Description of Included Studies**

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 where practices were randomized to usual care or
study intervention.\textsuperscript{119} One of the cohort studies were rated as fair quality,\textsuperscript{123} and one was rated as poor.\textsuperscript{124} The cross-sectional study was rated as good quality,\textsuperscript{125} the pre-post intervention study\textsuperscript{126} and quasi-experimental study\textsuperscript{127} were rated as poor quality, and the RCT as fair quality.\textsuperscript{119} All six studies were conducted in the United States.\textsuperscript{119,123-127} All studies 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; the sixth study used the PRIME-MD PHQ.\textsuperscript{127}

**Detailed Synthesis**

**Timing**

A good-quality cross-sectional study\textsuperscript{125} assessed 293 U.S. women at 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 at delivery, and 14 percent 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 at delivery, and 15 percent 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 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.\textsuperscript{123} 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). Obstetric 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\textsuperscript{127} sought to examine the 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 PHQ 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 (p<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. The quasi-experimental study design created potential for selection, detection, and performance biases.

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.\textsuperscript{126} 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.

A poor-quality prospective cohort study sought to determine whether universal depression screening during pregnancy and at 6 weeks postpartum affected rates of seeking treatment after recommendations for followup behavioral health assessments. The cohort consisted of 2,199 pregnant women who received obstetric care in a large multispecialty group practice in the United States. Postpartum depression was assessed using the EPDS. Patients scoring 9 or higher were alerted to their elevated score and encouraged by the obstetrician to seek a behavioral health provider. For patients who scored 14 or greater, the patient’s obstetrician or nurse assisted the patient during the visit to schedule a behavioral health appointment (unless the patient declined the referral). Of the 2,199 new obstetric patients screened during the universal screening program, 412 (18.7%) scored 9 or higher, and 102 (4.6%) scored 14 or higher. Of the 102 patients who scored 14 or higher, none followed the recommendations to be assessed by a behavioral health provider. Of the original cohort, 569 had progressed to the 6-week postpartum visit and had screening data available via a chart review. Of these 569 women, 28 (4.9%) had an EPDS of 14 or higher, and 5 (17.9%) had followed recommendations to seek care. There was no systematic analysis of factors affecting the probability of seeking additional care.

**Setting**

In a fair-quality RCT, 28 of 33 primary care practices completed a study in which they were randomized into two arms: usual care with training about postpartum depression and an active arm with more extensive training and implementation of the EPDS for screening of women between 5 and 12 weeks postpartum. Subjects in both arms completed the EPDS, but scores were not provided to the usual care sites. Initial followup within the intervention practices for those with an elevated EPDS was with a practice-administered PHQ-9. Women in the usual-care practices completed the EPDS and PHQ-9, which were submitted to a central study site instead of to the clinicians in the practice. Overall, there were 990 women in the usual care group and 1,353 in the intervention group. Among these, 255 (26%) in the usual care group and 399 (29%) in the intervention group had an EPDS ≥10 or PHQ-9 ≥10. Overall, women in the intervention group with elevated EPDS scores were more likely to be diagnosed with depression than those in the usual care group (66% vs. 41%; p=0.001). Similarly, those in the intervention group with elevated EPDS scores were more likely to receive medication (56% vs. 35%; p<0.001), counseling (20% vs. 11%; p=0.02), or both (60% vs. 37%; p<0.0001). These differences in treatment rates appear to be almost entirely due to differences in the initial detection of depression—rates of treatment were almost identical in the two groups among those women who did receive a diagnosis of depression (89.7% of women with a diagnosis of depression received
medication and counseling in the usual care group vs. 90.7% in the intervention group). Although these rates of treatment are substantially higher than those reported in other settings, the study design, where usual care practices were blinded to EPDS scores, precludes drawing any direct inferences about whether provision of screening, diagnosis, and treatment within the same practice setting improves the likelihood of an appropriate response to an abnormal screening result.
Discussion

Key Findings and Strength of Evidence

In this comparative effectiveness review (CER), we reviewed 40 unique studies represented by 45 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 specifically 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. There were also four studies for the Beck Depression Inventory (BDI), but different versions of the test were used. There were two studies of the “two-question” screen, both of which found a sensitivity of 100 percent if the response to either question were “yes,” but with markedly lower specificities than other tests (45.5% and 65.7%).

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

<table>
<thead>
<tr>
<th>Screening Test</th>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Risk of Bias</th>
<th>Consistency</th>
<th>Directness</th>
<th>Precision</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Test Performance (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antenatal Risk Questionnaire</td>
<td>Sensitivity</td>
<td>1 (276)</td>
<td>High</td>
<td>NA</td>
<td>Direct</td>
<td>Imprecise</td>
<td></td>
<td>Low SOE 78.1% (65.0–88.7%)</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (276)</td>
<td>High</td>
<td>NA</td>
<td>Direct</td>
<td>Imprecise</td>
<td></td>
<td>Low SOE 47.1% (40.3–59.9%)</td>
</tr>
<tr>
<td>BDI</td>
<td>Sensitivity</td>
<td>2 (1,151)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
<td>Imprecise</td>
<td></td>
<td>Low SOE 80–90% (approximate range of point estimates at most commonly used thresholds)</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>2 (1,151)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
<td>Precise</td>
<td></td>
<td>Low SOE 80–90% (approximate range of point estimates at most commonly used thresholds)</td>
</tr>
<tr>
<td>BDI-II</td>
<td>Sensitivity</td>
<td>2 (650)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
<td>Imprecise</td>
<td></td>
<td>Low SOE 75–90% (approximate range of point estimates at most commonly used thresholds)</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>2 (650)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
<td>Precise</td>
<td></td>
<td>Low SOE 80–90% (approximate range of point estimates at most commonly used thresholds)</td>
</tr>
<tr>
<td>EPDS</td>
<td>Sensitivity</td>
<td>11 (3,456)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
<td>Imprecise</td>
<td></td>
<td>Moderate SOE 80–90% (approximate range of point estimates at most commonly used thresholds)</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>11 (3,456)</td>
<td>Medium</td>
<td>Consistent</td>
<td>Direct</td>
<td>Precise</td>
<td></td>
<td>Moderate SOE 80–90% (approximate range of point estimates at most commonly used thresholds)</td>
</tr>
</tbody>
</table>
Table 14. Strength-of-evidence domains for test characteristics of screening tests for postpartum depression (continued)

<table>
<thead>
<tr>
<th>Screening Test</th>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Test Performance (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
</tr>
<tr>
<td>HRSD-17</td>
<td>Sensitivity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td>HRSD-21</td>
<td>Sensitivity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td>Leverton</td>
<td>Sensitivity</td>
<td>1 (617)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>Specificity</td>
<td>1 (617)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td>PDSS</td>
<td>Sensitivity</td>
<td>4 (903)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>4 (903)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
</tbody>
</table>
Table 14. Strength-of-evidence domains for test characteristics of screening tests for postpartum depression (continued)

<table>
<thead>
<tr>
<th>Screening Tests</th>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Test Performance (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>Sensitivity</td>
<td>1 (506)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>1 (506)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td>Two-Question Screen</td>
<td>Sensitivity</td>
<td>2 (600)</td>
<td>Low</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>2 (600)</td>
<td>Low</td>
<td>Consistent</td>
</tr>
</tbody>
</table>

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-9 = Patient Health Questionnaire; SOE = strength of evidence

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. In the 2005 AHRQ evidence report,2,3 the estimated point prevalence of major depression at various points in the first 12 months after delivery was in the 4–8 percent range; the prevalence in the majority of studies included in this review was in the 10–20 percent range. There are approximately 4,000,000 deliveries annually in the United States.128 Table 15 shows the effect of prevalence, sensitivity, and specificity on the estimated annual number of true positives, false positives, and false negatives if all postpartum women are screened once during the postpartum period. It is clear from these numbers that, although a 10-percent difference in either sensitivity or specificity may appear relatively small, there are significant differences in both the number of missed diagnoses and the number of false positives across this range. Even at a relatively high prevalence, decreasing specificity from 90 to 80 percent results in over 300,000 additional false positives annually. Even if false-positive results have no individual harms, this would represent either a substantial strain on existing resources for evaluation of women with possible depression or require a substantial investment in additional resources. (The implications of this tradeoff if screening is repeated throughout the postpartum year are discussed below).
Table 15. 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

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

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. Another study compared prevalence of suicidal ideation based on the EPDS to another scale, the MOODS-SR. Prevalence of suicidal ideation was approximately twice as high on the EPDS, but the investigators did not formally compare agreement between the two or compare either to a reference standard. In this study, not surprisingly, suicidal ideation on the EPDS was significantly associated with a subsequent diagnosis of major depression.

KQ 2. Effect of Individual 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.

The majority of these factors are consistent predictors of postpartum depression in earlier studies included in other reviews, and it is possible that including older studies would have raised the overall strength of evidence based on greater consistency or precision. However, given that there is evidence that temporal trends in the methods used to classify subjects as depressed or nondepressed affect study results, this is not at all certain.
The purpose of our review of this literature was ultimately not to assess whether a given risk factor is or is not associated with postpartum depression, but whether screening women with the risk factor results in better test performance—and even more importantly—better clinical outcomes compared to screening women without the risk factor. We did not identify any studies (even observational studies) that made this direct comparison. This means that, even including additional studies, the strength of evidence that screening based on risk factors might improve performance would be moderate at best.

The potential clinical impact of better estimates of the association between a given risk factor (or group of factors) and postpartum depression is dependent not only on the strength of the association (as measured by the relative risk or odds ratio), but also on the baseline risk of postpartum depression and the prevalence of the risk factor—a common risk factor might result in a clinically significant increase in absolute risk even at low to moderate levels of increased relative risk. Given an estimate of the relative risk, the prevalence of the risk factor, and the incidence of postpartum depression, it is possible to estimate the absolute difference in incidence between those with and without the risk factor. This in turn would allow an estimation of how test characteristics, particularly positive and negative predictive value, would change if screening were conditional on the presence or absence of the risk factor. However, this estimate would again be indirect at best, and would require confirmation from more direct studies.

We did not identify any studies meeting our inclusion criteria that evaluated a risk prediction instrument (analogous to the use of risk prediction instruments such as the Gail model for breast cancer risk, which is used as a tool for deciding on both timing of screening and type of test). Multivariate predictive models can be characterized in terms of sensitivity and specificity. For screening for postpartum depression, a predictive model could be used to identify women with a higher pretest probability of depression (which in turn would improve positive predictive value). Alternatively, the results of a screening instrument could be incorporated into the model itself.

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

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal Demographics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>3 (5,578)</td>
<td>Medium Inconsistent Direct Imprecise Insufficient</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>2 (4,757)</td>
<td>Medium Inconsistent Direct Imprecise Insufficient</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>1 (4,245)</td>
<td>Medium NA Direct Imprecise Insufficient</td>
<td></td>
</tr>
<tr>
<td>Employment status (unemployed vs. employed)</td>
<td>1 (363)</td>
<td>High NA Direct Imprecise</td>
<td>Low SOE for increased risk of postpartum depression in unemployed mothers OR, 2.8 (1.1–4.9)</td>
</tr>
</tbody>
</table>
Table 16. Strength-of-evidence domains for associations with patient characteristics and risk of postpartum depression (continued)

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
<td>Directness</td>
</tr>
<tr>
<td>Obstetric History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parity</td>
<td>2 (4,998)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Preterm/low birthweight infant</td>
<td>2 (4,711)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Smoking</td>
<td>2 (4,998)</td>
<td>Medium</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>1 (4,348)</td>
<td>Medium</td>
<td>NA</td>
</tr>
<tr>
<td>General Medical History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor health status/chronic illness</td>
<td>2 (4,993)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Obesity</td>
<td>1 (598)</td>
<td>Medium</td>
<td>NA</td>
</tr>
<tr>
<td>Psychiatric History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of perinatal depression</td>
<td>2 (1,082)</td>
<td>High</td>
<td>Consistent</td>
</tr>
<tr>
<td>History of depression</td>
<td>5 (2,057)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>History of premenstrual dysphoric disorder</td>
<td>1 (210)</td>
<td>Medium</td>
<td>NA</td>
</tr>
<tr>
<td>Any psychiatric diagnosis</td>
<td>2 (1,075)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Anxiety</td>
<td>2 (1,305)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Personality (vulnerable/ neuroticism)</td>
<td>2 (685)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Relationship/ Social Support</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status (single/no relationship)</td>
<td>3 (5,803)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Poor relationship quality</td>
<td>5 (6,101)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Poor social support</td>
<td>4 (1,830)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
</tbody>
</table>

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 directly comparing screening instrument performance across settings or type of provider (Table 17).

Table 17. Strength-of-evidence domains for the effect of varying timing on screening for postpartum depression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
</tr>
<tr>
<td>Delivery to 8 weeks vs. 8 weeks to 6 months</td>
<td>1 (534)</td>
<td>High</td>
<td>NA</td>
</tr>
<tr>
<td>Delivery vs. 6 weeks</td>
<td>1 (113)</td>
<td>High</td>
<td>NA</td>
</tr>
</tbody>
</table>

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 to moderate strength of evidence, with the strength of evidence from consistent results weakened because of poor to fair study quality and imprecise estimates), and 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) (Table 18). Parental stress as measured by the Parental Stress Inventory (PSI) or the PSI-Short Form (PSI-SF) did not improve with screening and treatment of depressive symptoms in a poor-quality quasi-experimental study, two fair-quality RCTs, and one good-quality RCT (low strength of evidence due to mostly poor to fair study quality and lack of precision), despite improvement in depressive symptoms with screening and treatment in all four studies. These results are consistent with a 2008 systematic review of the association between treatment of maternal depression and child outcomes, which concluded, “Based on [ten] studies, there is some evidence of associations between successful treatment of parents’ depression and improvement in children’s symptoms and functioning, but treatment of postpartum depression may not be sufficient for improving cognitive development, attachment, and temperament in infants and toddlers.”16

It is important to note that the lack of improvement observed in the PSI in the studies in our review does not necessarily mean that screening and treatment for depression are ineffective in improving important aspects of the mother–infant relationship. Other possible explanations include (1) interventions that are effective in reducing depressive symptoms, when used alone, may not be sufficient to improve parenting, particularly in settings where parental stress or
dysfunction is already high, (2) if sample sizes were based on change in response to a depression scale, and the PSI is not as sensitive to changes secondary to improved depressive changes, then the studies may have been underpowered to detect a difference in the PSI, (3) the impact of effective depression treatment on parenting takes longer to become evident than changes in depressive symptoms themselves, and (4) effective depression treatment could improve aspects of the mother–infant relationship not measured by the PSI. If part of the reason for emphasizing screening and treatment of depression in the postpartum period (compared to other points in adulthood) is to improve the mother–infant relationship, and longer term outcomes in the child, then identifying appropriate measures of this relationship—and appropriate study designs to measure them—needs to be a key research priority.

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 after adjusting for prescreen infant health status, but this difference was no longer significant by 12 months; it is unclear whether these visits represented inappropriate utilization. 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

<table>
<thead>
<tr>
<th>Benefits/ Harms</th>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Domains Pertaining to SOE</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Risk of Bias</td>
<td>Consistency</td>
</tr>
<tr>
<td>Benefits</td>
<td>Depressive symptoms</td>
<td>5 (8,071)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td></td>
<td>Mental health score (SF-12)</td>
<td>1 (2,579)</td>
<td>Low</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Parental stress</td>
<td>4 (5,567)</td>
<td>Medium</td>
<td>Consistent</td>
</tr>
<tr>
<td>Harms</td>
<td>Unscheduled doctor visits for infant</td>
<td>1 (462)</td>
<td>Medium</td>
<td>NA</td>
</tr>
</tbody>
</table>

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 KQs were low, ranging from 0 to 30 percent. Differences in country, setting, population characteristics, screening instrument, and timing precluded synthesis across studies. Three 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), a second found a much smaller difference when comparing prenatal (33%) with postpartum (27%) screening (p=not statistically significant [NS]), and a third poor-quality study found higher rates of followup among postpartum women (17.9%) compared with antepartum women (0%) (p=NS).

Although we did not identify any studies that directly addressed potential differences in appropriate followup based on setting or provider, there is some intriguing indirect evidence that practice characteristics may be very important. Reported followup and treatment rates among women with a positive screening test or clinical suspicion of depression were substantially higher in a study where screening, diagnosis, and treatment all occurred within an integrated primary care practice than were observed in other studies where positive screening results required referral for further diagnosis and treatment.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Number of Studies (Subjects)</th>
<th>Risk of Bias</th>
<th>Consistency</th>
<th>Directness</th>
<th>Precision</th>
<th>SOE and Magnitude of Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenatal vs. postpartum</td>
<td>3 (1,263)</td>
<td>Medium</td>
<td>Inconsistent</td>
<td>Direct</td>
<td>Imprecise</td>
<td>Low SOE for Higher rates of referral/diagnosis prenatally</td>
</tr>
<tr>
<td>Delivery vs. postpartum</td>
<td>1 (230)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>Imprecise</td>
<td>Low SOE for Higher rates of referral/diagnosis during delivery admission</td>
</tr>
</tbody>
</table>

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

Findings in Relation to What Is Already Known

Our review focused on studies published subsequent to the 2005 AHRQ evidence review on perinatal depression. Key findings of the 2005 AHRQ review included:

- Patient characteristics in the identified studies did not reflect the diversity of the U.S. population of pregnant and postpartum women.
- There was a lack of precision for estimates of test characteristics, particularly for test sensitivity.
- There were widely overlapping confidence intervals for estimates, precluding indirect comparison across tests.
- Relatively few studies were identified that directly compared results of multiple screening instruments.
There was overall better sensitivity of screening instruments for the detection of major depression compared with major and minor depression combined.

No studies compared screening with no screening.

Recommendations in the review 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
- Directly comparing different screening instruments within studies
- Conducting studies that evaluate a broader range of timing
- Designing studies that compare screening with no screening

Our findings in this review were broadly consistent with the 2005 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); studies directly comparing different screening instruments, and studies comparing screening with no screening. However, the overall strength of the evidence base is not much better now than it was in 2005. 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 studies that considered the recommendations of the 2005 report in their design have not yet been published.

A 2009 report for the Institute of Medicine, while not a formal systematic review, broadly reviewed the evidence for screening and treatment of depression in parents, including postpartum depression, and drew heavily on topic-specific systematic reviews, including the 2005 AHRQ report. The IOM report emphasized the consistent observational evidence of an association between parental depression and adverse short- and long-term outcomes in children. Specific summary conclusions regarding screening included:

Although there is evidence for effectiveness of screening, it is most effective when systems are in place to ensure adequate followup and treatment (similar to the USPSTF assessment).

There is a lack of data on the effect of screening in the primary care setting on parental function, barriers to utilization of services, or the two-generation impact of depression.

Although effective screening tools are available, patients are only identified as parents during the prenatal period.

A variety of programs have focused on screening mothers during routine pregnancy and postpartum clinical visits and other child health visits. These approaches provide opportunities to identify individuals who are at a higher risk for depression, provide education and support, assess parental function, and link child development screening with maternal depression screening (although the report reached no conclusions about effectiveness).

Studies have examined screening for depression in parents—particularly mothers—in existing community programs (e.g., early Head Start, those serving homeless women, substance use disorder treatment, home visitation), where individuals who are at higher risk of depression are seen. Although these settings and programs offer opportunities to reach parents and their children at greater risk for depression, screening is not routine (and, again, evidence on overall effectiveness is limited).

Little information is available in either public or private settings about the complex process of implementing a systematic approach to maternal or paternal depression screening and followup, including time, resources needed, workforce and training competency and capacity, and the impact of engagement and education of depressed parents on themselves as well as their children.
The findings of our review are consistent with these other reviews as well as with the USPSTF review and recommendations for screening in adults: there are reasonably consistent estimates for the sensitivity and specificity of available screening instruments, and there is evidence that screening and treatment can improve depressive symptoms; but the effectiveness of screening is dependent on the availability of systematic resources for managing patients with positive screening results, with the task force explicitly recommending screening only if such resources are available (with a “C” recommendation against screening if they are not). We identified many of the same uncertainties noted in these previous reviews, including a lack of evidence that there are no harms associated with screening (as opposed to not reporting of harms), a lack of evidence that screening and treatment for depression directly improves maternal–infant functioning, and a lack of evidence on the optimal screening interval.

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 qualitatively assessed the applicability of the included studies to a broader U.S. perspective. Many included studies recruited populations whose demographics differed considerably from 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%). Qualitatively, results in terms of test performance, risk factors, outcomes, or receipt of appropriate services did not consistently differ between U.S.–based studies compared to those conducted in other countries. 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. In addition, given large differences between countries in educational systems, social support resources, and other factors that contribute to longer term developmental outcomes, the extent to which effective treatment of postpartum depression may influence these longer term outcomes may differ as well. Many studies had highly selected samples due to high rates of nonresponse 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 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 to 13), and descriptions of testing protocols were not specific enough to inform routine clinical care. As discussed elsewhere, the choice of cutpoint has significant implications for clinical outcomes, at both individual patient level and health system level. 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 to no screening, and, although they suggest an improvement in depressive symptoms, there are limited data on other maternal or infant health outcomes. Lastly, there is limited information on paternal outcomes.

It is also worth noting that the single U.S.–based study that demonstrated high rates of receipt of appropriate services and significant reductions with screening\textsuperscript{119} did so within the context of family physician practices where integrated screening, diagnosis, and treatment services were available. However, the most recent available data suggest that, in the United States, family physicians account for less than 10 percent of prenatal visits (with presumably a similar proportion for postpartum visits)\textsuperscript{132} and less than 20 percent of nonacute visits for children under 4 years of age.\textsuperscript{133} If the majority of care for women or infants is being provided in settings where integration of screening with appropriate mental health diagnostic and treatment services is not available, then these results are not broadly applicable without a major change in current patterns of obstetric and pediatric care, which is unlikely in the short term.

**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. 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.”\textsuperscript{5} 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. 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.\textsuperscript{119} which is consistent with the USPSTF review.

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. No matter what methods are used to ensure appropriate followup, the resources required are directly dependent on the test characteristics of the screening test, as discussed throughout this report. A small decline in specificity can result in a large absolute increase in the number of positive results, most of which
will be false positives. 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.

**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 tradeoff between false-positive and false-negative test results (Table 20). Given estimates of the point prevalence of depression of 3–7 percent in the postpartum period 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 to generate preliminary estimates of these tradeoffs 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—a strategy frequently used in other contexts (for example, use of nontreponemal tests for syphilis, followed by more specific treponemal antigen tests in positive patients). One possible option would be to use the two-question screen, which had a reported sensitivity in two studies of 100 percent with specificities of 44 and 65 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.

Table 20 shows the expected number of false positives and false negatives for a one-time screen with (a) one of seven screening tests alone or (b) using one of the tests only after a positive response to one of the two questions that make up the two-question screen. This analysis assumes a prevalence of postpartum depression of 5.8 percent at 2 months postpartum (the highest point prevalence estimate in the 2005 AHRQ report) and universal screening. The estimates shown are the result of 10,000 simulations using randomly selected point estimates for sensitivity and specificity from the studies reviewed for KQ 1. Serial testing has a small effect on false-negative rates but substantially decreases false-positive rates for all tests. This decrease is most dramatic for tests with lower specificity. (Confidence intervals for the estimates are not shown in Table 20, but there is considerable overlap between tests—this table should not be used to draw inferences for between-test comparisons.) As noted above, even if a false-positive result does not have any significant impact on health outcomes at the individual level, evaluating and ruling out depression in women with false-positive screening results increases the workload for existing service providers and creates the need for additional resources, which may not be readily available, particularly for providers caring for vulnerable populations where resources are already constrained.
Table 20. Estimated annual number of true positives, false positives, and false negatives in the United States from screening with “single test” versus “serial tests”

<table>
<thead>
<tr>
<th>Screening Test</th>
<th>True Positives</th>
<th>False Positives</th>
<th>False Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single Test</td>
<td>Serial Tests</td>
<td>Single Test</td>
</tr>
<tr>
<td>Two questions</td>
<td>229,040</td>
<td>–</td>
<td>2,085,920</td>
</tr>
<tr>
<td>ANQR</td>
<td>171,320</td>
<td>169,000</td>
<td>1,973,640</td>
</tr>
<tr>
<td>BDI</td>
<td>121,840</td>
<td>120,080</td>
<td>212,680</td>
</tr>
<tr>
<td>EPDS</td>
<td>185,280</td>
<td>182,840</td>
<td>352,240</td>
</tr>
<tr>
<td>LQ</td>
<td>217,960</td>
<td>215,160</td>
<td>324,520</td>
</tr>
<tr>
<td>PDSS</td>
<td>181,520</td>
<td>179,120</td>
<td>606,520</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>196,040</td>
<td>193,480</td>
<td>617,040</td>
</tr>
</tbody>
</table>

ANQR = Antenatal Risk Questionnaire; BDI = Beck Depression Inventory; EPDS = Edinburgh Postnatal Depression Scale; LQ = Leverton Questionnaire; PDSS = Postpartum Depression Screening Scale; PHQ = Patient Health Questionnaire-9

*a“Single test” refers to results if indicated test used alone; “serial tests” refers to use of indicated test only if response to one of “two questions” is positive.

A better understanding of the tradeoffs between harms and benefits would help to identify the optimal test and strategy. As an example, Figure 9 presents the results of a microsimulation comparing no screening, screening with the EPDS alone, screening with the Postpartum Depression Screening Scale (PDSS) alone, screening 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 specificity were randomly drawn from the distributions described in each study described in KQ 1. (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.135) 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 tradeoff 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 proportion 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 acceptability 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 9 shows the following: serial testing is almost always favored over a single test; there is minimal difference between the EPDS and PDSS given the available evidence; and, even with serial testing, there is likely to be a high number of false positives associated with screening. If additional evidence were available on the clinical harms (as well as costs) associated with a false-positive result, making a recommendation for or against screening (either screening of any type or with a specific test) would be much easier.
Figure 9. Acceptability curve for tradeoff between false positives ("costs") and treated depression ("effectiveness") at different thresholds for false positives/treated depression ratio ("willingness-to-pay")

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.

In the studies reviewed, followup rates for women with positive screening results were uniformly low. The impact of these low followup rates on the overall effectiveness of screening is unclear. The false-positive rate of most of the screening instruments studied is high. Therefore, if the majority of women who did not get further evaluation after screening represented women who were truly not depressed, then the overall effectiveness of screening might not be substantially worsened. On the other hand, if women with true-positive 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. This may be particularly problematic for those providing services for low-income populations, where resources for mothers and infants are already under considerable strain. 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 (although, given the inconsistency of the results and findings related to setting, this may be related primarily to greater ease of access of referral services around the time of delivery). The risk for postpartum depression appears to continue at least through the first 12 months after delivery.\textsuperscript{2,3} The best estimate for cumulative incidence from birth to 12 months in the 2005 AHRQ report was approximately 30 percent (roughly 3% per month). This ongoing risk suggests that screening throughout the postpartum period might be necessary to maximize the detection of depression, particularly if doing so is necessary to optimize parenting.

However, as screening frequency increases, so does the likelihood of false-positive results for both individuals and the population—this effect has been clearly been demonstrated with cancer screening models.\textsuperscript{136} Estimating the impact of different screening frequencies in a cohort of postpartum women is difficult, even with an estimate of incidence, since the point prevalence at any given time is a function of (a) incidence, (b) the duration of symptoms/condition, and (c) the proportion of symptomatic women who will be diagnosed in between screening intervals. For illustration, we can make assumptions favorable to screening, including (1) all of the new cases of depression will remain undiagnosed if screening is not performed, (2) none of the new cases will spontaneously remit in the absence of screening, (3) all women with true-positive results receive treatment, and (d) since women with false-positive results at one screening test will still be at risk for developing depression, they will be rescreened at the next scheduled time.

During each screening round, some women will have true-positive results and be removed from the cohort. At the next screening round, the total number of women with depression will be the sum of new cases among nondepressed women (true negatives and false positives in the previous round) and cases that were missed (false negatives) in the previous round. Table 21 shows the expected cumulative number of true positives, false positives, and false negatives in a cohort of 4 million women (the approximate number of deliveries in the United States annually) if screening is performed at a postpartum visit at 6 to 8 weeks, with subsequent screens during well-child visits at 3, 6, 9 and 12 months. We used the best estimates for prevalence at 6 to 8 weeks (8%), and cumulative incidence (approximately 30% at 12 months, or 3% per month) from the 2005 AHRQ report, at three different levels of sensitivity and specificity consistent with the ranges found in our review.
Table 21. Estimated number of true positives, false positives, and false negatives with screening at postpartum and well-child visits

<table>
<thead>
<tr>
<th>Time Since Birth</th>
<th>Sensitivity 90% Specificity 80%</th>
<th>Sensitivity 85% Specificity 85%</th>
<th>Sensitivity 80% Specificity 90%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2 months</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positives</td>
<td>291,600</td>
<td>275,400</td>
<td>259,200</td>
</tr>
<tr>
<td>False Positives</td>
<td>735,200</td>
<td>551,400</td>
<td>367,600</td>
</tr>
<tr>
<td>False Negatives</td>
<td>32,400</td>
<td>48,600</td>
<td>64,800</td>
</tr>
<tr>
<td><strong>3 months</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positives</td>
<td>130,366</td>
<td>136,894</td>
<td>141,801</td>
</tr>
<tr>
<td>False Positives</td>
<td>712,710</td>
<td>534,532</td>
<td>356,355</td>
</tr>
<tr>
<td>False Negatives</td>
<td>14,485</td>
<td>24,158</td>
<td>35,450</td>
</tr>
<tr>
<td><strong>6 months</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positives</td>
<td>298,456</td>
<td>290,097</td>
<td>282,066</td>
</tr>
<tr>
<td>False Positives</td>
<td>649,283</td>
<td>486,962</td>
<td>324,642</td>
</tr>
<tr>
<td>False Negatives</td>
<td>33,162</td>
<td>51,194</td>
<td>70,517</td>
</tr>
<tr>
<td><strong>9 months</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positives</td>
<td>289,864</td>
<td>289,088</td>
<td>287,541</td>
</tr>
<tr>
<td>False Positives</td>
<td>591,501</td>
<td>443,626</td>
<td>295,751</td>
</tr>
<tr>
<td>False Negatives</td>
<td>32,207</td>
<td>51,015</td>
<td>71,885</td>
</tr>
<tr>
<td><strong>12 months</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positives</td>
<td>265,865</td>
<td>267,082</td>
<td>268,067</td>
</tr>
<tr>
<td>False Positives</td>
<td>538,862</td>
<td>404,146</td>
<td>269,431</td>
</tr>
<tr>
<td>False Negatives</td>
<td>29,541</td>
<td>47,132</td>
<td>67,017</td>
</tr>
<tr>
<td><strong>Cumulative</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1,276,151</td>
<td>1,258,560</td>
<td>1,238,675</td>
</tr>
<tr>
<td>%</td>
<td>31.9%</td>
<td>31.5%</td>
<td>31.0%</td>
</tr>
</tbody>
</table>

Even at a specificity of 90 percent, repeated testing results in a 40 percent chance of having at least one false-positive test result in the first postpartum year; at lower levels of specificity, well over half of all women would have at least one false-positive result.

**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 is 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. As discussed above, inclusion of studies on risk factors for postpartum depression published prior to 2004 might have led to more precise estimates of the association, assuming no temporal trends in the use of specific diagnostic criteria, although it is unlikely that these earlier studies would have provided more direct evidence that screening based on the presence of risk factors results in different clinical outcomes.

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. 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 and the 2009 IOM report are still present and include the following:

- Patient characteristics in the applicable studies that 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 conducted in more diverse populations, additional studies are needed. This is particularly important given the need to increase the precision of estimates of test characteristics and more accurately determine the potential for variations in the prevalence of depression across diverse 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 formal screening to no screening or usual care; we identified only two fair-quality randomized controlled trials (RCT). Lack of evidence for 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 (a lack that was also noted in the USPSTF review of depression screening in the general adult population). Potential harms of a false-positive result at the individual level (or of a true-positive result when effective treatment is not available) include stigmatization and anxiety. Other than one study that reported a short-term increase in the number of unscheduled doctor visits in
infants of screened women (and where there was ambiguity about whether these visits were appropriate or not), we did not identify any studies that reported on outcomes for all women with positive results rather than limiting the reporting to only those women 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 weighing of harms versus benefits when deciding 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. Three studies of different design and different setting found no significant improvement in the PSI, a commonly used measure of parental stress among women screened and treated for postpartum depression, despite improvement in depressive symptoms. Whether this lack of change is an issue related to different levels of effectiveness of the interventions studies for depression and parenting, responsiveness of the specific measure used, or aspect of study design such as sample size, these results suggest that detection and treatment of depression alone may not be sufficient to lead to improved child outcomes. 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, for example, 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. This barrier was also specifically cited by the IOM in its review of depression in parents.4

**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,137 and there are two current funding opportunities from NIH specifically targeting mental health during pregnancy and the postpartum period138,139 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 not only is clinically meaningless (since noninteger values of number of deliveries have no interpretation) but also does not reflect the underlying distribution.

For many of the recommendations below, use of 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 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. 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 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; 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. These considerations should include 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 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.

- 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 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

- There was insufficient direct evidence to address the effect of timing, setting, or provider on test characteristics. It seems plausible that differences in clinical outcome relevant to timing, setting, or provider are more directly related to aspects of the process of screening, referral, and diagnosis rather 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 on appropriate targets for sensitivity and specificity.
- Parental stress should be included in studies of screening and treatment of maternal depression. Furthermore, the relationship between stress, depression, and other important outcomes should be carefully explored.
- 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, 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.
- 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 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 the variability in outcomes by setting, indirect comparisons across a small number of studies suggest that the receipt of appropriate services is much higher when screening, diagnosis, and treatment are provided by the same provider or practice, and depressive symptoms are substantially improved. 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


73. The World Bank. Country and lending groups. 


Abbreviations

AHRQ   Agency for Healthcare Research and Quality
BDI    Beck Depression Inventory
BDI-1A Beck Depression Inventory-1A
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-5  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
MADRS  Montgomery Asberg Depression Rating Scale
MOODS-SR Mood Spectrum Self-Report
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
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
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSI</td>
<td>Parenting Stress Index</td>
</tr>
<tr>
<td>PSI-SF</td>
<td>Parenting Stress Index-Short Form</td>
</tr>
<tr>
<td>QALY</td>
<td>quality-adjusted life year</td>
</tr>
<tr>
<td>QUADAS-2</td>
<td>QUality Assessment of Diagnostic Accuracy Studies-2</td>
</tr>
<tr>
<td>RCT</td>
<td>randomized controlled trial</td>
</tr>
<tr>
<td>RDC</td>
<td>Research Diagnostic Criteria</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operating characteristic</td>
</tr>
<tr>
<td>RR</td>
<td>relative risk</td>
</tr>
<tr>
<td>SADS</td>
<td>Schedule for Affective Disorders and Schizophrenia</td>
</tr>
<tr>
<td>SCAN</td>
<td>Schedules for Clinical Assessment in Neuropsychiatry</td>
</tr>
<tr>
<td>SCID</td>
<td>Structured Clinical Interview for Depression</td>
</tr>
<tr>
<td>SF-36</td>
<td>Medical Outcomes Study 36-Item Short-Form Health Survey</td>
</tr>
<tr>
<td>SPI</td>
<td>Goldberg's Standardized Psychiatric Interview</td>
</tr>
<tr>
<td>TEP</td>
<td>Technical Expert Panel</td>
</tr>
<tr>
<td>TOO</td>
<td>Task Order Officer</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>USPSTF</td>
<td>U.S. Preventive Services Task Force</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>Zung SDS</td>
<td>Zung Self-Rating Depression Scale</td>
</tr>
</tbody>
</table>
## Appendix A. Exact Search Strings

### PubMed® Search Strategy (July 24, 2012)

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#2</td>
<td>Depression[Mesh] OR Depressive Disorder[Mesh] OR depression[tiab]</td>
</tr>
<tr>
<td>#3</td>
<td>#1 AND #2</td>
</tr>
<tr>
<td>#4</td>
<td>postpartum period/psychology[mesh] OR depression, postpartum[mesh]</td>
</tr>
<tr>
<td>#5</td>
<td>#3 OR #4</td>
</tr>
<tr>
<td>#7</td>
<td>#5 AND #6</td>
</tr>
<tr>
<td>#8</td>
<td>#7 NOT (Editorial[ptyp] OR Letter[ptyp] OR Case Reports[ptyp] OR Comment[ptyp]) NOT (animals[mh] NOT humans[mh])</td>
</tr>
<tr>
<td>#9</td>
<td>#8 Limits: English, 2004 - present</td>
</tr>
</tbody>
</table>

### PsycINFO® Search Strategy (July 24, 2012)

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>((DE &quot;Prenatal Care&quot;) OR (DE &quot;Pregnancy&quot; OR DE &quot;Adolescent Pregnancy&quot;)) OR (DE &quot;Birth&quot; OR DE &quot;Natural Childbirth&quot; OR DE &quot;Premature Birth&quot;) OR TI ( prenatal OR perinatal OR postnatal OR pregnancy OR pregnant OR postpartum OR post-partum ) OR AB ( prenatal OR perinatal OR postnatal OR pregnancy OR pregnant OR postpartum OR post-partum )</td>
</tr>
<tr>
<td>S2</td>
<td>(DE &quot;Depression (Emotion)&quot;) OR (DE &quot;Major Depression&quot; OR DE &quot;Anacritic Depression&quot; OR DE &quot;Dysthymic Disorder&quot; OR DE &quot;Endogenous Depression&quot; OR DE &quot;Postpartum Depression&quot; OR DE &quot;Reactive Depression&quot; OR DE &quot;Recurrent Depression&quot; OR DE &quot;Treatment Resistant Depression&quot;) OR TI depression OR AB depression</td>
</tr>
<tr>
<td>S3</td>
<td>S1 AND S2</td>
</tr>
<tr>
<td>S4</td>
<td>DE &quot;Postpartum Depression&quot; OR DE &quot;Postpartum Psychosis&quot;</td>
</tr>
<tr>
<td>S5</td>
<td>S3 OR S4</td>
</tr>
<tr>
<td>S6</td>
<td>DE &quot;Screening&quot; OR DE &quot;Screening Tests&quot; OR DE &quot;Psychological Screening Inventory&quot; OR DE &quot;Rating Scales&quot; OR DE &quot;Inventories&quot; OR DE &quot;Psychological Assessment&quot; OR DE &quot;Psychodiagnosis&quot; OR DE &quot;Psychodiagnostic Interview&quot; OR DE &quot;Questionnaires&quot; OR DE &quot;General...&quot;</td>
</tr>
<tr>
<td>Set #</td>
<td>Terms</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>#1</strong></td>
<td>Health Questionnaire&quot; OR ((DE &quot;Beck Depression Inventory&quot;) OR (DE &quot;Zungs Self Rating Depression Scale&quot;)) OR TI ( questionnaire OR questionnaires OR screening OR screen OR scale OR instrument OR instruments OR EPDS OR “Edinburgh postnatal depression&quot; OR BDI OR &quot;beck depression inventory&quot; OR PDSS OR “Postpartum Depression Screening Scale&quot; OR BPDS OR “Bromley Postnatal Depression Scale&quot; OR LQ OR “Leverton Questionnaire&quot; OR CES-D OR “Center for Epidemiologic Studies Depression Scale&quot; OR HADS OR “Hospital Anxiety and Depression Scale&quot; OR PHQ-9 OR “Patient Health Questionnaire-9&quot; OR “Zung SDS&quot; OR “Zung Self-Rating Depression Scale&quot; OR HRSD OR “Hamilton Rating Scale for Depression&quot; OR PDPRI-R OR “Postpartum Depression Predictors Inventory-Revised&quot; OR GHQ-D OR “General Health Questionnaire&quot; OR MADRS OR “Montgomery Asberg Depression Rating Scale&quot; ) OR AB ( questionnaire OR questionnaires OR screening OR screen OR scale OR instrument OR instruments OR EPDS OR “Edinburgh postnatal depression&quot; OR BDI OR &quot;beck depression inventory&quot; OR PDSS OR “Postpartum Depression Screening Scale&quot; OR BPDS OR “Bromley Postnatal Depression Scale&quot; OR LQ OR “Leverton Questionnaire&quot; OR CES-D OR “Center for Epidemiologic Studies Depression Scale&quot; OR HADS OR “Hospital Anxiety and Depression Scale&quot; OR PHQ-9 OR “Patient Health Questionnaire-9&quot; OR “Zung SDS&quot; OR “Zung Self-Rating Depression Scale&quot; OR HRSD OR “Hamilton Rating Scale for Depression&quot; OR PDPRI-R OR “Postpartum Depression Predictors Inventory-Revised&quot; OR GHQ-D OR “General Health Questionnaire&quot; OR MADRS OR “Montgomery Asberg Depression Rating Scale&quot; OR “generalized contentment scale&quot; OR “patient health questionnaire-2&quot; OR &quot;phq-2&quot; OR “primary care evaluation of mental disorders patient health questionnaire&quot; OR &quot;prime-md phq&quot; )</td>
</tr>
<tr>
<td><strong>S7</strong></td>
<td>S5 AND S6</td>
</tr>
<tr>
<td><strong>S9</strong></td>
<td>S7 NOT S8</td>
</tr>
<tr>
<td><strong>S10</strong></td>
<td>S9, Limits: - Publication Year from: 2004-; Publication Type: All Journals; Language: English; Population Group: Human</td>
</tr>
</tbody>
</table>

**Embase® Search Strategy (July 24, 2012)**
Platform: Embase.com

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>'obstetric care'/exp OR 'pregnancy'/exp OR 'puerperal disorder'/exp OR prenatal:ab,ti OR perinatal:ab,ti OR postnatal:ab,ti OR pregnancy:ab,ti OR pregnant:ab,ti OR postpartum:ab,ti</td>
</tr>
<tr>
<td>#2</td>
<td>'depression'/exp OR depression:ab,ti</td>
</tr>
<tr>
<td>#3</td>
<td>#1 AND #2</td>
</tr>
<tr>
<td>#4</td>
<td>'puerperal depression'/exp</td>
</tr>
<tr>
<td>#5</td>
<td>#3 OR #4</td>
</tr>
<tr>
<td>#6</td>
<td>'puerperal depression'/exp dm_di OR 'screening'/exp OR 'questionnaire'/exp OR 'interview'/exp OR 'Edinburgh Postnatal Depression Scale'/exp OR 'Beck Depression Inventory'/exp OR 'Center for Epidemiological Studies Depression Scale'/exp OR 'Hospital Anxiety and Depression Scale'/exp OR 'General Health Questionnaire'/exp OR 'Montgomery Asberg Depression Rating Scale'/exp OR 'psychometry'/exp OR 'psychological rating scale'/exp OR questionnaire:ab,ti OR questionnaires:ab,ti OR screening:ab,ti OR screen:ab,ti OR scale:ab,ti OR instrument:ab,ti OR instruments:ab,ti OR EPDS:ab,ti OR &quot;Edinburgh postnatal depression&quot;:ab,ti OR BDI:ab,ti OR &quot;beck depression inventory&quot;:ab,ti OR PDSS:ab,ti OR &quot;Postpartum Depression Screening Scale&quot;:ab,ti OR BPDS:ab,ti OR &quot;Bromley Postnatal Depression Scale&quot;:ab,ti OR LQ:ab,ti OR “Leverton Questionnaire”:ab,ti OR &quot;CES D&quot;:ab,ti OR “Center for Epidemiologic Studies Depression Scale&quot;:ab,ti</td>
</tr>
</tbody>
</table>
Cochrane Search Strategy (July 24, 2012)
Platform: Wiley
Database searched: Cochrane Database of Systematic Reviews

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>MeSH descriptor Maternal Health Services explode all trees OR MeSH descriptor Pregnancy explode all trees OR MeSH descriptor Pregnant Women explode all trees OR MeSH descriptor Puerperal Disorders explode all trees OR prenatal:ti,ab OR perinatal:ti,ab OR postnatal:ti,ab OR pregnancy:ti,ab OR pregnant:ti,ab OR postpartum:ti,ab OR post-partum:ti,ab</td>
</tr>
<tr>
<td>#2</td>
<td>MeSH descriptor Depression explode all trees OR MeSH descriptor Depressive Disorder explode all trees OR depression:ti,ab</td>
</tr>
<tr>
<td>#3</td>
<td>#1 AND #2</td>
</tr>
<tr>
<td>#4</td>
<td>MeSH descriptor Postpartum Period explode all trees with qualifier: PX OR MeSH descriptor Depression, Postpartum explode all trees</td>
</tr>
<tr>
<td>#5</td>
<td>#3 OR #4</td>
</tr>
<tr>
<td>#6</td>
<td>MeSH descriptor Depression, Postpartum explode all trees with qualifier: DI OR MeSH descriptor Mass Screening explode all trees OR MeSH descriptor Questionnaires explode all trees OR MeSH descriptor Interviews as Topic explode all trees OR MeSH descriptor Psychometrics explode all trees OR MeSH descriptor Psychiatric Status Rating Scales explode all trees OR questionnaires:ti,ab OR screening:ti,ab OR screen:ti,ab OR scale:ti,ab OR instrument:ti,ab OR instruments:ti,ab OR EPDS:ti,ab OR &quot;Edinburgh postnatal depression&quot;:ti,ab OR BDI:ti,ab OR &quot;beck depression inventory&quot;:ti,ab OR PDSS:ti,ab OR &quot;Postpartum Depression Screening Scale&quot;:ti,ab OR BPDS:ti,ab OR &quot;Bromley Postnatal Depression Scale&quot;:ti,ab OR LQ:ti,ab OR &quot;Leverton Questionnaire&quot;:ti,ab OR CES-D:ti,ab OR “Center for Epidemiologic Studies Depression Scale”:ti,ab OR HADS:ti,ab OR “Hospital Anxiety and Depression Scale”:ti,ab OR PHQ-9:ti,ab OR “Patient Health Questionnaire-9”:ti,ab OR “Zung SDS”:ti,ab OR “Zung Self-Rating Depression Scale”:ti,ab OR HRSD:ti,ab OR “Hamilton Rating Scale for Depression”:ti,ab OR PDPI-R:ti,ab OR “Postpartum Depression Predictors Inventory-Revised”:ti,ab OR GHQ-D:ti,ab OR “General Health Questionnaire”:ti,ab OR MADRS:ti,ab OR &quot;Montgomery Asburg Depression Rating Scale&quot;:ti,ab OR “generalized contentment scale”:ti,ab OR “patient health questionnaire-2”:ti,ab OR “phq-2”:ti,ab OR “primary care evaluation of mental disorders patient health questionnaire”:ti,ab OR “prime-md phq”:ti,ab</td>
</tr>
<tr>
<td>#7</td>
<td>#5 AND #6</td>
</tr>
<tr>
<td>#8</td>
<td>#7, limit to Cochrane Database of Systematic Reviews, 2004 - present</td>
</tr>
</tbody>
</table>
Grey Literature Searches

ProQuest COS Conference Papers Index (July 24, 2012)

<table>
<thead>
<tr>
<th>Set #</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>all(&quot;Maternal Health Services&quot; OR Puerperal OR prenatal OR perinatal OR postnatal OR pregnancy OR pregnant OR postpartum OR post-partum)</td>
</tr>
<tr>
<td>#2</td>
<td>All(Depression)</td>
</tr>
<tr>
<td>#3</td>
<td>#1 AND #2</td>
</tr>
<tr>
<td>#4</td>
<td>all(diagnosis OR questionnaires OR interviews OR Psychometrics OR questionnaire OR screening OR screen OR scale OR instrument OR instruments OR EPDS OR &quot;Edinburgh postnatal depression&quot; OR BDI OR &quot;beck depression inventory&quot; OR PDSS OR &quot;Postpartum Depression Screening Scale&quot; OR BPDS OR &quot;Bromley Postnatal Depression Scale&quot; OR LQ OR &quot;Leverton Questionnaire&quot; OR CES-D OR “Center for Epidemiologic Studies Depression Scale” OR HADS OR “Hospital Anxiety and Depression Scale” OR PHQ-9 OR “Patient Health Questionnaire-9” OR “Zung SDS” OR “Zung Self-Rating Depression Scale” OR HRSD OR “Hamilton Rating Scale for Depression” OR PDPI-R OR “Postpartum Depression Predictors Inventory-Revised” OR GHQ-D OR “General Health Questionnaire” OR MADRS OR “Montgomery Asberg Depression Rating Scale” OR “generalized contentment scale” OR “patient health questionnaire-2” OR “phq-2” OR “primary care evaluation of mental disorders patient health questionnaire” OR “prime-md phq”)</td>
</tr>
<tr>
<td>#5</td>
<td>#3 AND #4</td>
</tr>
<tr>
<td>#6</td>
<td>#5, 2004 - present</td>
</tr>
</tbody>
</table>

ClinicalTrials.gov (August 22, 2012)

Search strategy: postpartum depression [ALL-FIELDS]

Total number of results: 117

WHO: International Clinical Trials Registry Platform Search Portal (August 22, 2012)

Search strategy: postpartum depression (standard search)

Total number of results: 93 records for 92 trials
## Appendix B. Data Abstraction Elements

### Study Characteristics

- **Study Identifiers**
  - Study Name or Acronym
  - Last name of first author
  - Publication year

- **Additional Articles Used in This Abstraction**

- **Study Dates**
  - Enrollment start (Mon and YYYY)
  - Enrollment end (Mon and YYYY)
  - Follow-up end (Mon and YYYY)

- **Study Sites**
  - Single Center, Multicenter, Unclear/Not reported
  - Number of sites

- **Geographic Location (Select all that apply)**
  - US, Canada, UK, Europe, S. America, C. America, Asia, Africa, Australia/NZ, Unclear/Not reported, Other (specify)

- **Study Design**
  - Prospective RCT
  - Prospective cohort
  - Retrospective cohort
  - Case-control
  - Cross-sectional
  - Pre-post-intervention
  - Other (specify)

- **Funding Source (Select all that apply)**
  - Government, Industry, Non-government/non-industry, Unclear/Not reported, Other (specify)

- **Setting (Select all that apply)**
  - Prenatal care, Hospital, Birthing Center, Home, Short-term postpartum follow-up, Well-child visit, Unclear/Not reported, Other (specify)

- **Provider (Select all that apply)**
  - Obstetricians, Family practitioners, Nurse-midwives, Mental health professionals, Lactation consultants, Social workers, Behavioral health specialists, Unclear/Not reported, Other (specify)

- **Enrollment Approach (Select all that apply)**
  - Consecutive patients, Convenience sample (not explicitly consecutive), Unclear/Not reported, Other (specify)

- **Study Inclusion and Exclusion Criteria**
  - Copy/paste inclusion and exclusion criteria as reported
  - Is the study entirely composed of participants with any of the following characteristics/conditions? If all participants fall into more than one category, select all that apply.
    - Specific race or ethnicity (specify)
    - Specific socioeconomic category (specify)
    - Specific parity (specify)
    - Specific cultural consideration (specify)
    - History of mood disorders
- All participants with normal perinatal outcome
- All participants with preterm perinatal outcome
- All participants with stillbirth perinatal outcome
- History of intimate partner violence
- None of the above

- Study Enrollment/ Study Completion
  - Number of participants (N) assessed for eligibility
  - N eligible
  - N enrolled/included
  - N completed follow-up (most distal time point of the primary outcome)
  - N analyzed for primary outcome

- Key Question Applicability (Select all that apply)
  - KQ 1: KQ 1a, KQ 1b
  - KQ 2: KQ 2a, KQ 2b
  - KQ 3: KQ 3a, KQ 3b, KQ 3c
  - KQ 4
  - KQ 5
  - KQ 6

- Comments

**Screening Intervention Characteristics** – Record the following elements for participants in Group 1, Group 2, Group 3, and Group 4 (as applicable)

- Screening Instrument
  - Edinburgh Postnatal Depression Scale (EPDS)
  - Beck Depression Inventory (BDI-IA)
  - Beck Depression Inventory (BDI-II)
  - Center for Epidemiologic Studies Depression Scale (CES-D)
  - General Health Questionnaire (GHQ-D)
  - Postpartum Depression Screening Scale (PDSS)
  - Hamilton Rating Scale for Depression (HRSD)
  - Zung Self-Rating Depression Scale (Zung SDS)
  - Patient Health Questionnaire-9 (PHQ-9)
  - Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MD PHQ)
  - Leverton Questionnaire (LQ)
  - Hospital Anxiety and Depression Scale (HADS)
  - Postpartum Depression Predictors Inventory-Revised (PDPI-R)
  - Montgomery Asberg Depression Rating Scale (MADRS)
  - Patient Health Questionnaire-2 (PHQ-2)
  - Bromley Postnatal Depression Scale (BPDS)
  - Generalized Contentment Scale (GCS)
  - Other (specify)

- Threshold for Positive Result

- Timing of Screening
  - Prenatal period
  - Perinatal (from admission for delivery to discharge)
  - Discharge to 8 weeks postpartum
  - >8 weeks to 12 months postpartum

- Setting
  - Prenatal care
Hospital
  - Birthing Center
  - Home
  - Short-term postpartum followup
  - Well-child visit
  - Unclear/Not reported
  - Other (specify)
- Provider
  - Obstetricians
  - Family practitioners
  - Nurse-midwives
  - Mental health professionals
  - Lactation consultants
  - Social workers
  - Behavioral health specialists
  - Unclear/Not reported
  - Other (specify)
- Intervention Descriptors
  - Describe the intervention received by participants in each group (Groups 1, 2, 3, and 4, as applicable).
- Diagnosis of Depression and Receipt of Services
  - N with a positive screening test
  - N referred for diagnostic evaluation
  - N who received a diagnostic evaluation
  - N with a true positive diagnosis
  - N with a diagnostic referral for treatment
  - N treated
- Specify the validated instrument used for diagnosis of depression

Baseline Population – Record the following elements for Total Population, Group 1, Group 2, Group 3, and Group 4 (as applicable)
- Number of participants in each group
- Gender
  - Female N
  - Male N
- Ethnicity
  - Hispanic or Latino
  - Not Hispanic or Latino
- Race
  - American Indian or Alaska Native
  - Asian
  - Black or African American
  - Native Hawaiian or other Pacific Islander
  - White
  - Multiracial
  - Other
- Age
  - Mean
  - Median
  - Standard Deviation
- Standard Error
- Min age
- Max age
- 25% IQR
- 75% IQR
- Categorical (specify distribution)

- Education
  - Mean
  - Median
  - Standard Deviation
  - Standard Error
  - IQR
  - Categorical (specify distribution)

- Language – Record N and % for the following:
  - English
  - Spanish
  - Other language (specify)

- Immigration
  - Native-born
    - N
    - %
  - Immigrant
    - N
    - %
  - Describe immigrant population

- Income (specify units)
  - Mean
  - Median
  - Standard Deviation
  - Standard Error
  - IQR
  - Categorical (specify distribution)

- Socioeconomic Status (specify units)
  - Mean
  - Median
  - Standard Deviation
  - Standard Error
  - IQR
  - Categorical (specify distribution)

- Social Support (specify units)
  - Mean
  - Median
  - Standard Deviation
  - Standard Error
  - IQR
  - Categorical (specify distribution)

- Marital Status – Record N and % for the following:
  - Married/Domestic Partnership
  - Unmarried
  - Other (specify)
• Perinatal Outcomes – Record N and % for the following:
  o Normal
  o Preterm
  o Stillbirth
  o Other (specify)
• Parity
  o Mean
  o Median
  o Standard Deviation
  o Standard Error
  o IQR
  o Categorical (specify distribution)
• History of Mood Disorders
  o N
  o %
• History of Intimate Partner Violence
  o N
  o %
• Breastfeeding
  o Yes: N, %
  o No: N, %
• Breastfeeding Duration
  o Mean
  o Median
  o Standard Deviation
  o Standard Error
  o IQR
  o Categorical (specify distribution)

Patient-Centered Outcomes
• Select the outcome reported on this form:
  o Receipt of appropriate diagnostic and treatment services for symptoms of depression
  o Scores on validated measures of maternal well-being and parenting
  o Breastfeeding
  o Scores on validated diagnostic instruments for depression
  o Health-related quality of life, based on validated measures
  o Maternal suicidal/infanticidal behaviors
  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
  o Scores on validated measures of stigmatization
  o Composite (report only if composed entirely of outcomes listed above)
  o No patient-centered outcomes of interest reported
• Additional details to describe outcome measure
• Time points to be abstracted (check all that apply)
• Delivery
  • Discharge to 8 weeks postpartum
  • Close to 6 months
  • Close to 1 year
  • Most distal time point after one year

  For each time point, record the following elements, as applicable:
  • Specify actual timing of outcome (include units)
  • Group: 1, 2, 3, 4
  • N Analyzed (enter UNK if unknown)
  • Unadjusted Result
    o Mean
    o Median
    o Mean within group change
    o Mean between group change
    o Number of patients with outcome
    o % of patients with outcome
    o Events/denominator
    o Odds ratio
    o Hazard ratio
    o Relative risk
    o Other (specify)
  • Unadjusted Result Variability
    o Standard Error (SE)
    o Standard Deviation (SD)
    o IQR
    o 95% CI
    o Other % CI (specify)
    o Other (specify)
  • Unadjusted Result, p-value between groups
  • Unadjusted Result, Reference group (for comparison between groups)
  • Adjusted Result
    o Mean
    o Median
    o Mean within group change
    o Mean between group change
    o Number of patients with outcome
    o % of patients with outcome
    o Events/denominator
    o Odds ratio
    o Hazard ratio
    o Relative risk
    o Other (specify)
  • Adjusted Result Variability
    o Standard Error (SE)
    o Standard Deviation (SD)
    o IQR
    o 95% CI
    o Other % CI (specify)
    o Other (specify)
  • Adjusted Result, p-value between groups
  • Adjusted Result, Reference group (for comparison between groups)
If adjusted data is recorded, indicate the adjustments applied

- Does the study report any subgroup analyses for this outcome? (Yes/No)
  - If Yes, describe the subgroup analyses and summarize results
- Comments

**Screening Instrument Performance**

- **Screening Test 1**
  - Edinburgh Postnatal Depression Scale (EPDS)
  - Beck Depression Inventory (BDI or BDI-IA)
  - Beck Depression Inventory (BDI-II)
  - Center for Epidemiologic Studies Depression Scale (CES-D)
  - General Health Questionnaire (GHQ-D)
  - Postpartum Depression Screening Scale (PDSS)
  - Hamilton Rating Scale for Depression (HRSD)
  - Zung Self-Rating Depression Scale (Zung SDS)
  - Patient Health Questionnaire-9 (PHQ-9)
  - Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MD PHQ)
  - Leverton Questionnaire (LQ)
  - Hospital Anxiety and Depression Scale (HADS)
  - Postpartum Depression Predictors Inventory-Revised (PDPI-R)
  - Montgomery Asberg Depression Rating Scale (MADRS)
  - Patient Health Questionnaire-2 (PHQ-2)
  - Bromley Postnatal Depression Scale (BPDS)
  - Generalized Contentment Scale (GCS)
  - Other (specify)

- **Screening Test 1 Positive Threshold**

- **Screening Test 2**
  - Edinburgh Postnatal Depression Scale (EPDS)
  - Beck Depression Inventory (BDI-IA)
  - Beck Depression Inventory (BDI-II)
  - Center for Epidemiologic Studies Depression Scale (CES-D)
  - General Health Questionnaire (GHQ-D)
  - Postpartum Depression Screening Scale (PDSS)
  - Hamilton Rating Scale for Depression (HRSD)
  - Zung Self-Rating Depression Scale (Zung SDS)
  - Patient Health Questionnaire-9 (PHQ-9)
  - Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MD PHQ)
  - Leverton Questionnaire (LQ)
  - Hospital Anxiety and Depression Scale (HADS)
  - Postpartum Depression Predictors Inventory-Revised (PDPI-R)
  - Montgomery Asberg Depression Rating Scale (MADRS)
  - Patient Health Questionnaire-2 (PHQ-2)
  - Bromley Postnatal Depression Scale (BPDS)
  - Generalized Contentment Scale (GCS)
  - Other (specify)
  - None

- **Screening Test 2 Positive Threshold**

- **Diagnostic Test**
• DSM-IV-TR criteria
• Research Diagnostic Criteria (RDC)
• Bedford College Checklist
• International Classification of Diseases (ICD)
• Structured Clinical Interview for Depression (SCID)
• Diagnostic Interview Schedule (DIS)
• Schedule for Affective Disorders and Schizophrenia (SADS)
• Goldberg's Standardized Psychiatric Interview (SPI)
• Other (specify)

• Diagnostic Test Positive Threshold
• Briefly describe the definition of postpartum depression used for each screening tool.
• List any other comparisons reported between different thresholds.
• Does this data represent a predictive model or algorithm? (Yes/No)
  o If Yes:
    ▪ Describe the model/algorithm.
    ▪ Capture the data for the model/algorithm in the tables below or following text box.
• Sensitivity/Specificity Data – Record the following elements for Total Population, Group 1, Group 2, Group 3, and Group 4 (as applicable)
  o Participant Data
    ▪ Number of participants who received screening test 1
    ▪ Number of participants who refused screening test 1
    ▪ Number of participants with positive screening test 1
    ▪ Number of participants with negative screening test 1
    ▪ Number of participants who received screening test 2
    ▪ Number of participants who refused screening test 2
    ▪ Number of participants with positive screening test 2
    ▪ Number of participants with negative screening test 2
    ▪ Number of participants who received the diagnostic test
    ▪ Number of participants who refused the diagnostic test
    ▪ Disease prevalence (N of participants)
    ▪ Disease prevalence (% of participants)
  o Screening Tool Results (recorded separately for screening tool 1 and screening tool 2)
    ▪ True positive (N)
    ▪ True negative (N)
    ▪ False positive (N)
    ▪ False negative (N)
    ▪ Indeterminate or technically inadequate results (N)
    ▪ Sensitivity (%)
    ▪ Sensitivity (Standard deviation)
    ▪ Sensitivity (Confidence interval range)
      ▪ 95% CI
      ▪ Other (specify)
    ▪ Specificity (%)
    ▪ Specificity (Standard deviation)
    ▪ Specificity (Confidence interval range)
      ▪ 95% CI
      ▪ Other (specify)
    ▪ Positive predictive value (%)
• Positive predictive value (Standard deviation)
• Positive predictive value (Confidence interval range)
  • 95% CI
  • Other (specify)
• Negative predictive value (%)
• Negative predictive value (Standard deviation)
• Negative predictive value (Confidence interval range)
  • 95% CI
  • Other (specify)

  Enter any pertinent information that cannot be captured in the tables above.

• Additional Questions
  • Were both the screening test and diagnostic test done on all subjects? (Yes, No, or Unclear/Not reported)
  • What was the time interval between the screening test and the diagnostic test?
  • Was the screening test interpreted in a blinded fashion without knowledge of results of other diagnostic tests or clinical history and risk factors? (Yes, No, or Unclear/Not reported)
  • Was the diagnostic test interpreted in a blinded fashion without knowledge of results of other diagnostic tests or clinical history and risk factors? (Yes, No, or Unclear/Not reported)
  • Describe any paternal outcomes reported.

Quality
• Did the study present clinical outcomes? (Yes/No)
  • If Yes, select the study type: RCT, Cohort or Pre-post, Case-control, Cross sectional
  • If RCT, select Yes/No/Unclear for each of the following questions:
    • Selection Bias
      • Was the allocation sequence generated adequately (e.g., random number table, computer-generated randomization)?
      • Was the allocation of treatment adequately concealed (e.g., pharmacy-controlled randomization or use of sequentially numbered sealed envelopes)?
      • Were participants analyzed within the groups they were originally assigned to?
      • Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches?
    • Performance Bias
      • Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?
      • Did the study maintain fidelity to the intervention protocol?
    • Attrition Bias
      • If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)?
    • Detection Bias
      • In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time
period between the intervention/exposure and outcome different for cases and controls?
• Were the outcome assessors blinded to the intervention or exposure status of participants?
• Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants?
• Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants?

- **Reporting Bias**
  • Were the potential outcomes prespecified by the researchers? Are all prespecified outcomes reported?

  - **If Cohort or Pre-post, select Yes/No/Unclear for each of the following questions:**
    - **Selection Bias**
      • Were participants analyzed within the groups they were originally assigned to?
      • Did the study apply inclusion/exclusion criteria uniformly to all comparison groups?
      • Did the strategy for recruiting participants into the study differ across study groups?
      • Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches?
    - **Performance Bias**
      • Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?
      • Did the study maintain fidelity to the intervention protocol?
    - **Attrition Bias**
      • If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)?

  - **Detection Bias**
    • In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time period between the intervention/exposure and outcome different for cases and controls?
    • Were the outcome assessors blinded to the intervention or exposure status of participants?
    • Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants?
    • Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants?
    • Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?

  - **Reporting Bias**
    • Were the potential outcomes prespecified by the researchers? Are all prespecified outcomes reported?

  - **If Case-control, select Yes/No/Unclear for each of the following questions:**
- **Selection Bias**
  - Were cases and controls selected appropriately (e.g., appropriate diagnostic criteria or definitions, equal application of exclusion criteria to case and controls, sampling not influenced by exposure status)?
  - Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches?

- **Performance Bias**
  - Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?
  - Did the study maintain fidelity to the intervention protocol?

- **Attrition Bias**
  - If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)?

- **Detection Bias**
  - In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time period between the intervention/exposure and outcome different for cases and controls?
  - Were the outcome assessors blinded to the intervention or exposure status of participants?
  - Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants?
  - Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants?
  - Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?

- **Reporting Bias**
  - Were the potential outcomes prespecified by the researchers? Are all prespecified outcomes reported?

  o If Cross-sectional, select Yes/No/Unclear for each of the following questions:
    - **Selection Bias**
      - Did the study apply inclusion/exclusion criteria uniformly to all comparison groups?
      - Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches?
    - **Performance Bias**
      - Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?
    - **Attrition Bias**
      - If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)?
    - **Detection Bias**
• Were the outcome assessors blinded to the intervention or exposure status of participants?
• Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants?
• Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants?
• Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?
  ▪ Reporting Bias
  • Were the potential outcomes prespecified by the researchers? Are all prespecified outcomes reported?
  o Other Bias
  ▪ If applicable, describe any other concerns that may impact risk of bias.
  o Overall Study Rating (Good/Fair/Poor)
  ▪ Good (low risk of bias). These studies have the least bias, and the results are considered valid. These studies adhere to the commonly held concepts of high quality, including the following: a clear description of the population, setting, approaches, and comparison groups; appropriate measurement of outcomes; appropriate statistical and analytical methods and reporting; no reporting errors; a low dropout rate; and clear reporting of dropouts.
  ▪ Fair. These studies are susceptible to some bias, but not enough to invalidate the results. They do not meet all the criteria required for a rating of good quality because they have some deficiencies, but no flaw is likely to cause major bias. The study may be missing information, making it difficult to assess limitations and potential problems.
  ▪ Poor (high risk of bias). These studies have significant flaws that may have invalidated the results. They have serious errors in design, analysis, or reporting; large amounts of missing information; or discrepancies in reporting.
  ▪ If the study is rated as “Fair” or “Poor,” provide rationale.
• Did the study present diagnostic data? (Yes/No)
• If Yes, indicate Yes/No/Unclear for each of the following questions:
  o Signaling questions
    ▪ Patient Selection
      • Was a consecutive or random sample of patients enrolled?
      • Was a case-control design avoided?
      • Did the study avoid inappropriate exclusions?
    ▪ Index Test
      • Were the index test results interpreted without knowledge of the results of the reference standard?
      • If a threshold was used, was it pre-specified?
    ▪ Reference Standard
      • Is the reference standard likely to correctly classify the target condition?
      • Were the reference standard results interpreted without knowledge of the results of the index test?
    ▪ Flow & Timing
• Was there an appropriate interval between index test(s) and reference standard?
• Did all patients receive a reference standard?
• Did all patients receive the same reference standard?
• Were all patients included in the analysis?

  o Risk of bias
    ▪ Patient Selection
      • Could the selection of patients have introduced bias?
    ▪ Index Test
      • Could the conduct or interpretation of the index test have introduced bias?
    ▪ Reference Standard
      • Could the reference standard, its conduct or its interpretation have introduced bias?
    ▪ Flow & Timing
      • Could the patient flow have introduced bias?

  o Concerns regarding applicability
    ▪ Patient Selection
      • Are there concerns that the included patients do not match the review question?
    ▪ Index Test
      • Are there concerns that the index test, its conduct, or interpretation differ from the review question?
    ▪ Reference Standard
      • Are there concerns that the target condition as defined by the reference standard does not match the review question?

• Overall study rating
  o High risk of bias/ Low risk of bias/ Unclear
• Comments

Applicability – Use the PICOS format to identify specific issues, if any, which may limit the applicability of the study to this review.

  • Population (P)
    o Narrow eligibility criteria and exclusion of those with comorbidities
    o Large differences between demographics of study population and community patients
    o Narrow or unrepresentative severity, stage of illness, or comorbidities
    o Run-in period with high-exclusion rate for non-adherence or side effects
    o Event rates much higher or lower than observed in population-based studies

  • Intervention (I)
    o Doses or schedules not reflected in current practice
    o Intensity and delivery of behavioral interventions that may not be feasible for routine use
    o Monitoring practices or visit frequency not used in typical practice
    o Older versions of an intervention no longer in common use
    o Co-interventions that are likely to modify effectiveness of therapy
    o Highly selected intervention team or level of training/proficiency not widely available

  • Comparator (C)
    o Inadequate comparison therapy
• Use of substandard alternative therapy

- Outcomes (O)
  - Composite outcomes that mix outcomes of difference significance
  - Short-term or surrogate outcomes

- Setting (S)
  - Standards of care differ markedly from setting of interest
  - Specialty population or level of care differs from that seen in community

- Comments
Appendix C. Included Studies

Below is a list of all included studies in alphabetical order. Inset citations marked with an asterisk did not individually meet criteria for inclusion but were considered for supplemental information (e.g., methods data pertinent to an included study) for the articles they follow. Related articles (representing the same studies) are indicated with lettered superscripts.


Postpartum depression in women with epilepsy versus women without epilepsy. Epilepsy Behav. 2006;9:293-7.


Appendix D. Excluded Studies

All studies listed below were reviewed in their full-text version and excluded for the reason indicated in bold. Reasons for exclusion signify only the usefulness of the articles for this study and are not intended as criticisms of the articles.

**Full-Text Unavailable**

**Not a Full Publication (Abstract Only), or Not Original Peer-Reviewed Data**


Conde A, Figueiredo B and Bifulco A. Father's attachment style and psychological adjustment during


Eaton CL and Savola-Levin KJ. One year design and implementation of an integrated system for screening, diagnosis, and treatment of perinatal depression at a large HMO. Archives of Women's Mental Health 2011;14(SUPPL. 1):S64-S65.


Fisher J, Rowe H and Wynter K. What were we thinking! An innovative psycho-educational program to prevent common postpartum mental disorders in


Howard L, Chew-Graham CA, Tylee A, et al. The RESPOND trial: A randomized evaluation of antidepressants and support for women with postnatal


Le Donne M, Settineri S and Benvena S. Early postpartum (PP) alexithymia (AT) and depression(D): Relationship with serumtsh, thyroid hormones (TH) and thyroid autoantibodies (TAB). Thyroid 2011;21(SUPPL. 1):A47.

Legato MJ. The skewed sex distribution in affective disorders--a diagnostic, social, or biological problem?. Prog Brain Res 2010;186:159-66. PMID: 21094891.


Maccio EM and Pangburn JA. The case for investigating postpartum depression in lesbians and bisexual women. Women's Health Issues 2011;21(3):187-190.


McErlean RA, Dadds MR and Austin MP. A randomised controlled trial on intranasal oxytocin as an adjunct to interaction coaching to improve maternal bonding in women with mild postpartum depression. Biological Psychiatry 2011;69(9):134S.


Meltzer-Brody SE. Phenotyping, prevalence, and genetic signature of postpartum depression in the nesda GWAS study. Biological Psychiatry 2012;71(SUPPL. 1):12S.


Norwitz ER, Bellanger K, Funai EF, et al. Post-Traumatic stress disorder (PTSD) and panic disorder are independent risk factors for preterm birth, but not low birth weight or small-for-gestational age: The Yale 'Pink & Blue' Prospective cohort study. Reproductive Sciences 2010;17(3):352A.


Palmer AA. Development of a support program to address the mental health needs of women along the continuum of perinatal care. Archives of Women's Mental Health 2011;14(SUPPL. 1):S78.


Rabenda-Lacka KM. The prevalence and risk factors of postpartum depression in Poland. Archives of Women's Mental Health 2011;14(SUPPL. 1):S44.


Sit D. Seasonality of depression risk and suicidal symptoms in the postpartum period. Neuropsychopharmacology 2010;35 SUPPL. 1:S378.


Spinelli MG and Endicott J. Controlled clinical trial of antepartum interpersonal psychotherapy versus parenting education program at 3 NYC sites. Archives of Women's Mental Health 2011;14(SUPPL. 1):S49.


Stramrood CAI, Van Geenen M, Doornbos B, et al. PTSD and depression in partners of women with PE or PPROM. Reproductive Sciences 2010;17(3):205A.


Tareen A, Kothari C, Cherto M, et al. Smoking in perinatal women is associated with intimate partner victimization, housing challenges and history of...


White JS and Wisner KL. Co-location of psychiatric services within a home visiting agency. Archives of Women's Mental Health 2011;14(SUPPL. 1):S85.


**Not a Randomized Trial, or Not an Observational Study of Appropriate Design With Sample Size >100**


Not a Population of Interest


Armstrong K and Edwards H. The effects of exercise and social support on mothers reporting depressive


Dennis CL. Peer support for postpartum depression: volunteers' perceptions, recruitment strategies and training from a randomized controlled trial. Health Promot Int 2012 Mar 1[Epub ahead of print] PMID: 22388589.


Flynn HA, Sexton M, Ratliff S, et al. Comparative performance of the Edinburgh Postnatal Depression Scale and the Patient Health Questionnaire-9 in pregnant and postpartum women seeking psychiatric


Parker T, Maviglia MA, Lewis PT, et al. Psychological distress among Plains Indian mothers with children referred to screening for Fetal Alcohol


**Study Locations Did Not Include at Least One High-Income Economy**


Le HN, Perry DF and Ortiz G. The Postpartum Depression Screening Scale-Spanish version: examining the psychometric properties and


Dietz PM, Williams SB, Callaghan WM, et al. Clinically identified maternal depression before,


No Comparator of Interest


Logsdon MC, Usui WM and Nering M. Validation of Edinburgh postnatal depression scale for adolescent...


Miller RL, Pallant JF and Negri LM. Anxiety and stress in the postpartum: is there more to postnatal distress than depression?. BMC Psychiatry 2006;6:12. PMID: 16563155.


Pawar G, Wetzkcr C and Gjerdingen D. Prevalence of depressive symptoms in the immediate postpartum...


No Outcomes of Interest


Austin MP, Hadzi-Pavlovic D, Leader L, et al. Maternal trait anxiety, depression and life event stress in pregnancy: relationships with infant...


Bernazzani O, Conroy S, Marks MN, et al. Contextual Assessment of the Maternity Experience:


Chien LY, Tai CJ, Yeh MC. Domestic decision-making power, social support, and postpartum depression symptoms among immigrant and native women in Taiwan. Nurs Res 2012;61(2):103-10. PMID: 22307142.


Clemmens D, Driscoll JW and Beck CT. Postpartum depression as profiled through the depression screening scale. MCN Am J Matern Child Nurs 2004;29(3):180-5. PMID: 15123975.


Cornish AM, McMahon CA, Ungerer JA, et al. Postnatal depression and infant cognitive and motor...


Hayes BA, Muller R and Bradley BS. Perinatal depression: a randomized controlled trial of an


Horowitz JA. Community-based postpartum depression screening within the first month after


Matthey S, Ross-Hamid C. Repeat testing on the Edinburgh Depression Scale and the HADS-A in pregnancy: Differentiating between transient and


Sit D, Seltman H and Wisner KL. Seasonal effects on depression risk and suicidal symptoms in postpartum women. Depress Anxiety 2011;28(5):400-5. PMID: 21381158.


Small R, Lumley J, Donohue L, et al. Randomised controlled trial of midwife led debriefing to reduce


Tabuchi N and Shimada K. Maternal characteristics and feelings of distress resulting from the crying of
high-risk infants. Japan Journal of Nursing Science

Tatano Beck C, Gable RK, Sakala C, et al.
Postpartum Depressive Symptomatology: Results
from a Two-Stage US National Survey. Journal of

Tees MT, Harville EW, Xiong X, et al. Hurricane
Katrina-related maternal stress, maternal mental
health, and early infant temperament. Matern Child

Maternal lifetime history of depression and
depressive symptoms in the prenatal and early
postnatal period do not predict infant-mother
attachment quality in a large, population-based Dutch
PMID: 22191607.

Thio IM, Oakley Browne MA, Coverdale JH, et al.
Postnatal depressive symptoms go largely untreated:
a probability study in urban New Zealand. Soc
PMID: 16838090.

Thome M, Alder EM and Ramel A. A population-
based study of exclusive breastfeeding in Icelandic
women: is there a relationship with depressive
symptoms and parenting stress?. Int J Nurs Stud

Thompson JF, Roberts CL and Ellwood DA.
Emotional and physical health outcomes after
significant primary post-partum haemorrhage (PPH):
a multicentre cohort study. Aust N Z J Obstet

Tiwari A, Chan KL, Fong D, et al. The impact of
psychological abuse by an intimate partner on the
mental health of pregnant women. BJOG
2008;115(3):377-84. PMID: 18190375.

Tortajada S, Garcia-Gomez JM, Vicente J, et al.
Prediction of postpartum depression using multilayer
perceptrons and pruning. Methods Inf Med

Tsai R and Schaffir J. Effect of depot
medroxyprogesterone acetate on postpartum
depression. Contraception 2010;82(2):174-7. PMID:
20654759.

Tucker JN, Grzywacz JG, Leng I, et al. Return to
work, economic hardship, and women's postpartum
health. Women Health 2010;50(7):618-38. PMID:
21104566.

Tuohy A and McVey C. Experience of pregnancy
and delivery as predictors of postpartum depression.
Psychol Health Med 2008;13(1):43-7. PMID:
18066918.

Urquia ML, O'Campo PJ, Heaman MI, et al.
Experiences of violence before and during pregnancy
and adverse pregnancy outcomes: an analysis of the
Canadian Maternity Experiences Survey. BMC
Pregnancy Childbirth 2011;11:42. PMID: 21649909.

Urquia ML, O'Campo PJ, Heaman MI. Revisiting
the immigrant paradox in reproductive health: The roles
of duration of residence and ethnicity. Social Science
& Medicine 2012;74(10):1610-1621. PMID:
22464222.

v. Ballestrem CL, Strauss M, Kachele H.
Contribution to the epidemiology of postnatal
depression in Germany—implications for the

Recent intimate partner violence as a prenatal
predictor of maternal depression in the first year
postpartum among Latinas. Arch Womens Ment

van Bussel JC, Spitz B and Demyttenaere K. Anxiety
in pregnant and postpartum women. An exploratory
study of the role of maternal orientations. J Affect
Disord 2009;114(1-3):232-42. PMID: 18793805.

van Bussel JC, Spitz B and Demyttenaere K.
Depressive symptomatology in pregnant and
postpartum women. An exploratory study of the role
of maternal antenatal orientations. Arch Womens

van Bussel JC, Spitz B and Demyttenaere K. Three
self-report questionnaires of the early mother-to-
infant bond: reliability and validity of the Dutch
version of the MPAS, PBQ and MIBS. Arch
Womens Ment Health 2010;13(5):373-84. PMID:
20127128.

van Bussel JC, Spitz B and Demyttenaere K.
Women's mental health before, during, and after
pregnancy: a population-based controlled cohort

van de Pol G, van Brummen HJ, Bruinse HW, et al.
Is there an association between depressive and
urinary symptoms during and after pregnancy?. Int
Urogynecol J Pelvic Floor Dysfunct

Pregnancy-related pelvic girdle pain in the


Wenzel A, Haugen EN, Jackson LC, et al. Anxiety symptoms and disorders at eight weeks postpartum. J


Published Prior to January 1, 2004


Chabrol H, Teissedre F, Saint-Jean M, et al. Prevention and treatment of post-partum depression:


### Appendix E. Study Characteristics Table

#### Appendix Table E-1. Characteristics of included studies

<table>
<thead>
<tr>
<th>Article/Study/ Applicable KQ</th>
<th>Study Details</th>
<th>Participant Flow</th>
<th>Population Characteristics</th>
<th>Screening Characteristics</th>
<th>Outcomes Reported</th>
<th>Study Quality; Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akincigil, 2010&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Design: Prospective cohort Location: U.S. Setting: Hospital Funding: Government Provider: NR</td>
<td>Assessed: 4365 Eligible: 4365 Enrolled: 4365 Completed: 4365 Analyzed: 4348</td>
<td>Sex: Female (4348, 100%) Age distribution (see Note at right): N &lt;22 yr=1520 N 22-24=830 N 25-34=1998 N &gt;34=944 Ethnicity: Hispanic or Latino N=1165 Race: Black/ African American N=2065, White N=944, Other N=165 Special population: None</td>
<td>Screening tool(s): CIDI-SF Timing: Perinatal Diagnostic comparator: CIDI-SF</td>
<td>Performance characteristics Scores on diagnostic instruments for depression (DSM-IV criteria)</td>
<td>Test performance: High risk of bias Patient-centered outcomes: Fair Note: Numbers reported under “Age distribution” at left reflect error in paper (total 5292, which is &gt;4365 assessed and &gt; 4348 analyzed</td>
</tr>
<tr>
<td>Andersson, 2006&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Design: Prospective cohort Location: Europe Setting: Hospital Funding: NR Provider: Obstetricians, research nurses</td>
<td>Assessed: 720 Eligible: 720 Enrolled: 650 Completed: 650 Analyzed: 650</td>
<td>Sex: Female (650, 100%) Mean age: 29.5 (SD 4.5) Ethnicity: NR Race: NR Special population: None</td>
<td>Screening tool(s): PRIME-MD PHQ Timing: &gt;8 wk to 12 mo Diagnostic comparator: PRIME-MD CEG</td>
<td>Performance characteristics Scores on diagnostic instruments for depression</td>
<td>Test performance: Unclear risk of bias Patient-centered outcomes: Good</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>--------------------------</td>
<td>------------------</td>
<td>---------------------</td>
</tr>
</tbody>
</table>
| **Austin, 2010**<sup>1</sup>  
**KQ 1** | Design: Prospective cohort  
Location: UK  
Setting: Hospital  
Funding: Government  
Provider: NR | Assessed: NR  
Eligible: 2250  
Enrolled: 1549  
Completed: 300  
Analyzed: 300 | Sex: Female (1549, 100%)  
Mean age: 31.3 (SD 4.43)  
Ethnicity: NR  
Race: NR  
Special population: None | Screening tool(s): EPDS, Interval symptom question  
Timing: >8 wk to 12 mo  
Diagnostic comparator: CIDI | Performance characteristics | Test performance: High risk of bias |
| **Barnes, 2009**<sup>2</sup>  
**Home Start**  
**KQ 2** | Design: Cluster randomized  
Location: UK  
Setting: Prenatal care, home  
Funding: Non-government, non-industry  
Provider: Home volunteer visitors | Assessed: 1007  
Eligible: 527  
Enrolled: 389  
Completed: 250  
Analyzed: 250 | Sex: Female (250, 100%)  
Mean age: 28.9 (SD 5.8)  
Ethnicity: NR  
Race: White N=203  
Special population: SDI ≥9 | Screening tool(s): EPDS  
Timing: >8 wk to 12 mo  
Diagnostic comparator: SCID | Performance characteristics  
Scores on diagnostic instruments for depression | Test performance: Low risk of bias  
Patient-centered outcomes: Fair |
| **Beck, 2005**<sup>3</sup>  
**KQ 2** (See Note at right) | Design: Cross-sectional  
Location: U.S.  
Setting: Short-term postpartum followup  
Funding: Non-government, non-industry  
Provider: Mental health professionals | Assessed: NR  
Eligible: NR  
Enrolled: 150  
Completed: 150  
Analyzed: 150 | Sex: Female (150, 100%)  
Mean age: 25.75 (SD 5.66)  
Ethnicity: Hispanic or Latino N=150  
Race: NR  
Special population: Hispanic | Screening tool(s): PDSS  
Timing: Discharge to 8 wk  
Diagnostic comparator: Clinical interview | Performance characteristics  
Scores on diagnostic instruments for depression | Test performance: High risk of bias  
Patient-centered outcomes: Fair  
Note: Same population as Beck, 2005<sup>4</sup> |
| **Beck, 2005**<sup>4</sup>  
**KQ 1** (See Note at right) | Design: Cross-sectional  
Location: U.S.  
Setting: Short-term postpartum followup  
Funding: Non-government, non-industry  
Provider: Mental health professionals | Assessed: NR  
Eligible: NR  
Enrolled: 150  
Completed: 150  
Analyzed: 150 | Sex: Female (150, 100%)  
Mean age: 25.75 (SD 5.66)  
Ethnicity: Hispanic or Latino N=150  
Race: NR  
Special population: Hispanic | Screening tool(s): PDSS  
Timing: Discharge to 8 wk, >8 wk to 12 mo  
Diagnostic comparator: Clinical interview | Performance characteristics | Test performance: High risk of bias  
Note: Same population as Beck, 2005<sup>4</sup> |
<table>
<thead>
<tr>
<th>Article/Study/ Applicable KQ</th>
<th>Study Details</th>
<th>Participant Flow</th>
<th>Population Characteristics</th>
<th>Screening Characteristics</th>
<th>Outcomes Reported</th>
<th>Study Quality; Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloch, 2005&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Design: Prospective cohort Location: Israel Setting: Hospital, home Funding: Government Provider: Mental health professionals</td>
<td>Assessed: NR Eligible: 1800 Enrolled: 318 Completed: 244 Analyzed: 244</td>
<td>Sex: Female (1800, 100%) Mean age: 30.4 (SD 5.6) Ethnicity: NR Race: NR Special population: None</td>
<td>Screening tool(s): EPDS + risk factor questionnaire Timing: Perinatal Diagnostic comparator: SCID</td>
<td>Performance characteristics Scores on diagnostic instruments for depression</td>
<td>Test performance: High risk of bias Patient-centered outcomes: Fair</td>
</tr>
<tr>
<td>Burton, 2011&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Design: Cross-sectional Location: U.S. Setting: Hospital Funding: NR Provider: NR</td>
<td>Assessed: 293 Eligible: 37 Enrolled: 37 Completed: 37 Analyzed: 37</td>
<td>Sex: Female (37, 100%) Age distribution: N &lt;20=3 N 20-34=32 N ≥25=2 Ethnicity/Race: Hispanic or Latino N=29, Black/African American N=4, White N=3, Other N=1 Special population: None</td>
<td>Screening tool(s): EPDS Timing: Prenatal, Discharge to 8 wk, Perinatal (from admission for delivery to discharge) Diagnostic comparator: Diagnostic evaluation</td>
<td>Receipt of appropriate diagnostic/treatment services for depression Patient-centered outcomes: Good</td>
<td></td>
</tr>
<tr>
<td>Chaudron, 2010&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Design: Cross-sectional Location: U.S. Setting: Well-child visit Funding: Government Provider: Pediatricians</td>
<td>Assessed: 647 Eligible: 639 Enrolled: 385 Completed: 198 Analyzed: 198</td>
<td>Sex: Female (198, 100%) Mean age: 24.6 (SD 5.6) Ethnicity: Hispanic or Latino N=14 Race: Black/African American N=137, White N=34, Other N=25 Special population: Low income and urban women</td>
<td>Screening tool(s): EPDS, BDI-II, PDSS Timing: Discharge to 8 wk, &gt;8 wk to 12 mo Diagnostic comparator: Clinical interview</td>
<td>Performance characteristics</td>
<td>Test performance: Low risk of bias</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>-----------------------------</td>
<td>---------------------------</td>
<td>------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Chee, 2008&lt;sup&gt;12&lt;/sup&gt; KQ 2 (See Note at right)</td>
<td>Design: Prospective cohort Location: Asia Setting: Hospital, obstetrics clinic in tertiary hospital Funding: Industry Provider: Study researcher</td>
<td>Assessed: 724 Eligible: 687 Enrolled: 559 Completed: 484 Analyzed: 471</td>
<td>Sex: Female (471, 100%) Age distribution: N &lt;21=4 N 21-35=373 N &gt;35=94 Ethnicity: NR Race: Chinese N=233, Other N=238 Special population: None</td>
<td>Screening tool(s): EPDS Timing: Perinatal, &gt;8 wk to 12 mo Diagnostic comparator: SCID IV</td>
<td>Performance characteristics Scores on diagnostic instruments for depression</td>
<td>Test performance: High risk of bias Patient-centered outcomes: Fair Note: Same population as Chee, 2005&lt;sup&gt;13&lt;/sup&gt;</td>
</tr>
<tr>
<td>Chee, 2005&lt;sup&gt;13&lt;/sup&gt; KQ 2 (See Note at right)</td>
<td>Design: Prospective cohort Location: Asia Setting: Hospital Funding: Government Provider: Mental health professionals</td>
<td>Assessed: 724 Eligible: 599 Enrolled: 559 Completed: 278 Analyzed: 278</td>
<td>Sex: Female (278, 100%) Mean age: 31 (SD 4.7) Ethnicity: Not Hispanic or Latino N=278 Race: Chinese 47.2%, Other 52.8% Special population: Singaporean women during confinement</td>
<td>Screening tool(s): EPDS Timing: Prenatal, Discharge to 8 wk Diagnostic comparator: Clinical interview</td>
<td>Performance characteristics Scores on diagnostic instruments for depression</td>
<td>Test performance: High risk of bias Patient-centered outcomes: Fair Note: Same population as Chee, 2008&lt;sup&gt;12&lt;/sup&gt;</td>
</tr>
<tr>
<td>Clarke, 2008&lt;sup&gt;14&lt;/sup&gt; KQ 1</td>
<td>Design: Cross-sectional Location: Canada Setting: Hospital, short-term postpartum followup Funding: Government Provider: NR</td>
<td>Assessed: NR Eligible: NR Enrolled: 103 Completed: 103 Analyzed: 103</td>
<td>Sex: Female (103, 100%) Mean age: 23.8 (SD 4.7) Ethnicity: Not Hispanic or Latino N=103 Race/special population: Canada First Nations and Metis</td>
<td>Screening tool(s): EPDS, BDI-II, PDSS Timing: Discharge to 8 wk, &gt;8 wk to 12 mo Diagnostic comparator: Clinical interview</td>
<td>Performance characteristics</td>
<td>Test performance: High risk of bias</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------</td>
<td>------------------</td>
<td>---------------------------</td>
<td>--------------------------</td>
<td>------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Csatorndai, 2009&lt;sup&gt;16&lt;/sup&gt;</td>
<td>Design: Cross-sectional Location: Europe Setting: Short-term postpartum followup Funding: NR Provider: Nurse-midwives</td>
<td>Assessed: 1921 Eligible: 1741 Enrolled: 1552 Completed: 617 Analyzed: 617</td>
<td>Sex: Female (1552, 100%) Mean age: 27.8 (SD 4.5) Ethnicity: NR Race: NR Special population: None</td>
<td>Screening tool(s): BDI (1A), LQ Timing: Discharge to 8 wk Diagnostic comparator: Structured clinical interview (DSM-IV)</td>
<td>Performance characteristics</td>
<td>Test performance: Low risk of bias</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>-----------------------------</td>
<td>----------------------------</td>
<td>------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Flynn, 2006&lt;sup&gt;21&lt;/sup&gt;</td>
<td>Design: Pre-post-intervention</td>
<td>Assessed: 1298</td>
<td>Sex: Female (73, 100%)</td>
<td>Screening tool(s): EPDS</td>
<td>Receipt of appropriate diagnostic/treatment services for depression</td>
<td>Patient-centered outcomes: Poor</td>
</tr>
<tr>
<td>KQ 6</td>
<td>Location: U.S. Setting: Prenatal care, short-term postpartum followup</td>
<td>Eligible: NR</td>
<td>Mean age: MDD+: 28.7 (SD 5.4)</td>
<td>Timing: Prenatal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Funding: NR Provider: Obstetricians, nurses</td>
<td>Enrolled: 73</td>
<td>MDD-: 31.4 (SD 4.5)</td>
<td>Diagnostic comparator: SCID-DSM-IV (Mood Module)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Completed: NR</td>
<td>Ethnicity: Hispanic or Latino N=2, Not Hispanic or Latino N=71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analyzed: 73</td>
<td>Race: Asian N=8, Black/ African American N=6, White N=55, Other N=2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Special population: None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garcia-Esteve, 2008&lt;sup&gt;22&lt;/sup&gt;</td>
<td>Design: Cross-sectional</td>
<td>Assessed: 1201</td>
<td>Sex: Female (334, 100%)</td>
<td>Screening tool(s): EPDS</td>
<td>Performance characteristics</td>
<td>Test performance: High risk of bias</td>
</tr>
<tr>
<td>KQ 2</td>
<td>Location: Europe Setting: Short-term postpartum followup Funding: Government Provider: NR</td>
<td>Eligible: 412</td>
<td>Age distribution: N &lt;20=9</td>
<td>Timing: Discharge to 8 wk</td>
<td>Scores on diagnostic instruments for depression</td>
<td>Patient-centered outcomes: Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Completed: 334</td>
<td>N 26–35=257</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analyzed: 334</td>
<td>N &gt;35=44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ethnicity: NR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Race: NR Special population: None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gjerdingen, 2011&lt;sup&gt;23&lt;/sup&gt;</td>
<td>Design: Prospective cohort</td>
<td>Assessed: NR</td>
<td>Sex: Female (506, 100%)</td>
<td>Screening tool(s): PHQ-9</td>
<td>Performance characteristics</td>
<td>Test performance: Low risk of bias</td>
</tr>
<tr>
<td>KQ 1 (See Note at right)</td>
<td>Location: U.S. Setting: Well-child visit Funding: Government Provider: Participant</td>
<td>Eligible: 1556</td>
<td>Mean age: 29.1 (SD 6.2)</td>
<td>Timing: Discharge to 8 wk, &gt;8 wk to 12 mo</td>
<td>Notes: Same population as Gjerdingen, 2009&lt;sup&gt;24&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enrolled: 506</td>
<td>Ethnicity: NR</td>
<td>Diagnostic comparator: SCID</td>
<td>N analyzed (506) includes all subjects who were enrolled and completed baseline interview, not just those who completed study (472)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Completed: 472</td>
<td>Race: Asian N=34, Black/ African American N=89, White N=339, Multiracial N=17, Other N=27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analyzed: 506</td>
<td>Special population: None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>-----------------------------</td>
<td>--------------------------</td>
<td>-------------------</td>
<td>---------------------</td>
</tr>
</tbody>
</table>
| **Gjerdingen, 2009**<sup>24</sup>  
KQ 1  
(See Note at right) | Design: Prospective cohort  
Location: U.S.  
Setting: Well-child visit  
Funding: Government Provider: NR | Assessed: 1988  
Eligible: 1556  
Enrolled: 506  
Completed: 469  
Analyzed: 469 | Sex: Female (506, 100%)  
Mean age: 29.1 (SD 6.2)  
Ethnicity: NR  
Race: American Indian or Alaska Native N=7, Asian N=34, Black/African American N=89, White N=339, Multiracial N=17, Other N=6, Not reported N=14  
Special population: None | Screening tool(s): PHQ-9, PHQ-2, 2-question screen  
Timing: Discharge to 8 wk, >8 wk to 12 mo  
Diagnostic comparator: SCID | Performance characteristics | Test performance: Low risk of bias  
Notes: Same population as Gjerdingen, 2011<sup>23</sup> |
| **Glavin, 2010**<sup>25</sup>  
KQ 4 | Design: Prospective cohort (quasi-experimental)  
Location: Europe  
Setting: Home  
Funding: University Provider: NR | Assessed: 3111  
Eligible: 2508  
Enrolled: 2247  
Completed: 754  
Analyzed: 754 | Sex: Female (754, 100%)  
Mean age: 32.5 (SD 4.4)  
Ethnicity: NR  
Race: NR  
Special population: None | Screening tool(s): EPDS  
Timing: Discharge to 8 wk, >8 wk to 12 mo  
Diagnostic comparator: None | Maternal well-being/ parenting scores (Parenting Stress Index)  
Scores on diagnostic instruments for depression (EPDS ≥10 at 1 year by group) | Patient-centered outcomes: Poor |
| **Goodman, 2010**<sup>26</sup>  
KQ 6 | Design: Prospective cohort  
Location: U.S.  
Setting: Prenatal care, home  
Funding: Non-government, non-industry Provider: Participant | Assessed: 659  
Eligible: NR  
Enrolled: 525  
Completed: 491  
Analyzed: 299 | Sex: Female (299, 100%)  
Mean age: 31.6 (SD 5.35)  
Race/Ethnicity: Hispanic or Latino N=65, White N=193, Other N=81, Not reported N=2  
Special population: None | Screening tool(s): EPDS  
Timing: Prenatal, Discharge to 8 wk  
Diagnostic comparator: Documentation in medical records of diagnosis, referrals, treatment | Receipt of appropriate diagnostic/treatment services for depression | Patient-centered outcomes: Fair |
<table>
<thead>
<tr>
<th>Article/Study/ Applicable KQ</th>
<th>Study Details</th>
<th>Participant Flow</th>
<th>Population Characteristics</th>
<th>Screening Characteristics</th>
<th>Outcomes Reported</th>
<th>Study Quality; Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ji, 2011**</td>
<td>Design: Prospective cohort Location: U.S. Setting: NR Funding: Government Provider: NR</td>
<td>Assessed: NR Eligible: 708 Enrolled: 708 Completed: 534 Analyzed: 534</td>
<td>Sex: Female (534, 100%) Mean age: 33.1 (SD 5.1) Ethnicity: Hispanic or Latino N=16, Not Hispanic or Latino N=518 Race: American Indian or Alaska Native N=12, Asian N=12, Black/ African American N=51, White N=458, Multiracial N=1 Special population: None</td>
<td>Screening tool(s): EPDS, BDI, HRSD-17, HSRS-21 Timing: Prenatal, Discharge to 8 wk, &gt;8 wk to 12 mo Diagnostic comparator: SCID (Mood Module)</td>
<td>Performance characteristics</td>
<td>Test performance: High risk of bias</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>-----------------------------</td>
<td>--------------------------</td>
<td>------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Kersting, 2007&lt;sup&gt;44&lt;/sup&gt;</td>
<td>Design: Prospective cohort Location: Europe Setting: Dept. of Gynecology and Obstetrics, University of Muenster Funding: NR Provider: Multidisciplinary team</td>
<td>Assessed: NR Eligible: NR Enrolled: 127 Completed: 89 Analyzed: 127</td>
<td>Sex: Female (127, 100%) Mean age: 33.2 (SD 4.9) Ethnicity: NR Race: NR Special population: None</td>
<td>Screening tool(s): BDI-II Timing: Discharge to 8 wk, &gt;8 wk to 12 mo Diagnostic comparator: SCID</td>
<td>Performance characteristics Scores on diagnostic instruments for depression</td>
<td>Test performance: High risk of bias</td>
</tr>
<tr>
<td>Leung, 2011&lt;sup&gt;41&lt;/sup&gt;</td>
<td>Design: RCT Location: Asia Setting: Well-child visit, Maternal and Child Health Centers Funding: NR Provider: Nurse-midwives</td>
<td>Assessed: 1249 Eligible: 552 Enrolled: 462 Completed: 430 Analyzed: 333</td>
<td>Sex: Female (462, 100%) Mean age: NR Ethnicity: Not Hispanic or Latino (100%) Race: Asian (100%) Special population: Chinese</td>
<td>Screening tool(s): EPDS Timing: Discharge to 8 wk, &gt;8 wk to 12 mo Diagnostic comparator: None</td>
<td>Maternal well-being/ parenting scores (GHQ-12) Maternal well-being/ parenting scores (Parenting Stress Inventory Total, Parenting Stress Inventory-Parental Distress, Parenting Stress Inventory-Parent Child Dysfunctional Interaction, GHQ-12) Infant health system resource utilization (Number of doctor visits, number of hospitalizations)</td>
<td>Patient-centered outcomes: Fair</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>----------------------------</td>
<td>----------------------------</td>
<td>-------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Mann, 2012 Born in Bradford Study KQ 1</td>
<td>Design: Prospective cohort Location: UK Setting: Prenatal care, short-term postpartum followup, home (telephone) Funding: Non-government, non-industry Provider: Behavioral health specialists</td>
<td>Assessed: 268 Eligible: 261 Enrolled: 155 Completed: Phase 1=126, Phase 2=94 Analyzed: Phase 1=126, Phase 2=94</td>
<td>Sex: Female (152, 100%) Mean age: 27.4 (SD 5.8) Ethnicity: Not Hispanic or Latino (100%) Race (of 152 eligible): Asian N=48, Black/African American N=6, White N=86, Multiracial N=7, Other N=5 Special population: None</td>
<td>Screening tool(s): Case-finding questions Timing: Prenatal, Discharge to 8 wk Diagnostic comparator: DSM-IV-TR criteria</td>
<td>Performance characteristics</td>
<td>Test performance: Low risk of bias</td>
</tr>
<tr>
<td>Mauri, 2010 Perinatal Research and Screening Unit Study KQ 2 (See Note at right)</td>
<td>Design: Prospective cohort Location: Europe Setting: Hospital Funding: Government; non-profit, and industry Provider: Mental health professionals</td>
<td>Assessed: 2138 Eligible: 2138 Enrolled: 1066 Completed: 500 Analyzed: 500</td>
<td>Sex: Female (1066, 100%) Mean age: 32.27 (SD 3.95) Ethnicity: Not Hispanic or Latino (100%) Race: NR Special population: Italian</td>
<td>Screening tool(s): EPDS Timing: Perinatal, Prenatal, Discharge to 8 wk, &gt;8 wk to 12 mo Diagnostic comparator: SCID</td>
<td>Performance characteristics Receipt of appropriate diagnostic/treatment services for depression</td>
<td>Test performance: High risk of bias Patient-centered outcomes: Fair Note: Same population as Mauri, 2012</td>
</tr>
<tr>
<td>Mauri, 2012 Perinatal Research and Screening Unit Study KQ 1 (See Note at right)</td>
<td>Design: Prospective cohort Location: Europe Setting: Hospital Funding: Government; non-profit, and industry Provider: Mental health professionals</td>
<td>Assessed: 2138 Eligible: 2138 Enrolled: 1066 Completed: 500 Analyzed: 500</td>
<td>Sex: Female (1066, 100%) Mean age: 32.3 (SD 3.9) Ethnicity: Not Hispanic or Latino (100%) Race: NR Special population: None</td>
<td>Screening tool(s): EPDS, MOODS-SR Timing: Prenatal, Perinatal, Discharge to 8 wk, &gt;8 wk to 12 mo Diagnostic comparator: SCID</td>
<td>Performance characteristics</td>
<td>Test performance: High risk of bias Patient-centered outcomes: Fair Note: Same population as Mauri, 2010</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------</td>
<td>-----------------</td>
<td>---------------------------</td>
<td>--------------------------</td>
<td>-------------------</td>
<td>---------------------</td>
</tr>
</tbody>
</table>
| Morrell, 2009** KQ 4       | Design: RCT  
Location: UK  
Setting: Well-child visit  
Funding: Government  
Provider: Health visitor | Assessed: NR  
Eligible: 7649  
Enrolled: 4084  
Completed: 418  
Analyzed: 418 | Sex: Female (418, 100%)  
Mean age: 30.9 (SD 5.4)  
Ethnicity: NR  
Race: White N=390  
Special population: None | Screening tool(s): EPDS  
Timing: Discharge to 8 wk, >8 wk to 12 mo  
Diagnostic comparator: None | Scores on diagnostic instruments for depression  
HRQOL (SF-12 PCS)  
Parental Depression (PSI-SF)  
Maternal well-being/ parenting scores (SF-12 MCS) | Patient-centered outcomes: Good |
| Navarro, 2007*** KQ 1      | Design: Cross-sectional  
Location: Europe  
Setting: Hospital, Obstetrics and Gynaecology Unit of teaching hospital  
Funding: NR  
Provider: Mental health professionals | Assessed: NR  
Eligible: 1453  
Enrolled: 1453  
Completed: 405  
Analyzed: 405 | Sex: Female (1453, 100%)  
Age distribution: N ≤18=18  
N 19-34=1044  
N ≥35=391  
Ethnicity: NR  
Race: NR  
Special population: None | Screening tool(s): EPDS, GHQ-12  
Timing: Discharge to 8 wk  
Diagnostic comparator: SCID | Performance characteristics | Test performance: High risk of bias |
| Pereira, 2010*** KQ 1      | Design: Prospective cohort  
Location: Europe  
Setting: Prenatal care, home  
Funding: Government  
Provider: Mental health professionals | Assessed: NR  
Eligible: 486  
Enrolled: 486  
Completed: 452  
Analyzed: 452 | Sex: Female (452, 100%)  
Mean age: 30.47 (SD 4.304)  
Ethnicity: NR  
Race: NR  
Special population: Normal perinatal outcome | Screening tool(s): BDI-II, PDSS  
Timing: >8 wk to 12 mo  
Diagnostic comparator: DIDS and OPCRIT | Performance characteristics | Test performance: Low risk of bias |
<table>
<thead>
<tr>
<th>Article/Study/ Applicable KQ</th>
<th>Study Details</th>
<th>Participant Flow</th>
<th>Population Characteristics</th>
<th>Screening Characteristics</th>
<th>Outcomes Reported</th>
<th>Study Quality; Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rowan, 2012</td>
<td>Design: Prospective cohort Location: U.S. Setting: Prenatal care, hospital Funding: Non-government, non-industry Provider: Obstetricians</td>
<td>Assessed: NR Eligible: NR Enrolled: 2199 Completed: 569 Analyzed: 569</td>
<td>Sex: Female (100%) Mean age: NR Ethnicity: NR Race: NR Special population: None</td>
<td>Screening tool(s): EPDS Timing: Discharge to 8 wk Diagnostic comparator: None</td>
<td>Receipt of appropriate diagnostic/treatment services for depression</td>
<td>Patient-centered outcomes: Poor</td>
</tr>
<tr>
<td>Article/Study/ Applicable KQ</td>
<td>Study Details</td>
<td>Participant Flow</td>
<td>Population Characteristics</td>
<td>Screening Characteristics</td>
<td>Outcomes Reported</td>
<td>Study Quality; Notes</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>----------------------------</td>
<td>--------------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Yawn, 2012</strong>&lt;sup&gt;TRIPPD&lt;/sup&gt; &lt;br&gt;KQ 4, KQ 6</td>
<td>Design: RCT &lt;br&gt;Location: U.S. &lt;br&gt;Setting: Family medicine research network practices &lt;br&gt;Funding: Government Provider: Family practitioners, nurses</td>
<td>Assessed: NR &lt;br&gt;Eligible: 2398 &lt;br&gt;Enrolled: 2343 &lt;br&gt;Completed: 1689 &lt;br&gt;Analysed: 397</td>
<td>Sex: Female (2343, 100%) &lt;br&gt;Mean age: &lt;br&gt;Intervention group: 26.1 (5.4) &lt;br&gt;Usual care group: 26.7 (5.6) &lt;br&gt;Ethnicity: Intervention group: Hispanic or Latino 18% &lt;br&gt;Usual care group: Hispanic or Latino 14% &lt;br&gt;Race: Black/ African American 18% &lt;br&gt;Special population: None</td>
<td>Screening tool(s): EPDS, PHQ-9 &lt;br&gt;Timing: Discharge to 8 wk, &gt;8 wk to 12 mo &lt;br&gt;Diagnostic comparator: MD clinical impression plus positive PHQ-9</td>
<td>Receipt of appropriate diagnostic/ treatment services for depression</td>
<td>Patient-centered outcomes: Fair</td>
</tr>
<tr>
<td><strong>Yonkers, 2009</strong>&lt;sup&gt;Healthy Start&lt;/sup&gt; &lt;br&gt;KQ 6</td>
<td>Design: Quasi-experimental (pre-post with two cohorts for comparators) &lt;br&gt;Location: U.S. &lt;br&gt;Setting: Hospital, Healthy Start Programs &lt;br&gt;Funding: Government Provider: Social workers</td>
<td>Assessed: NR &lt;br&gt;Eligible: 1336 &lt;br&gt;Enrolled: 1336 &lt;br&gt;Completed: NR &lt;br&gt;Analysed: 1336</td>
<td>Sex: Female (1336, 100%) &lt;br&gt;Mean age: 24.7 (SD 5.8) &lt;br&gt;Ethnicity: Hispanic or Latino N=665, Not Hispanic or Latino N=671 &lt;br&gt;Race: Black/ African American N=454, White N=176, Other N=40 &lt;br&gt;Special population: None</td>
<td>Screening tool(s): PRIME-MD PHQ &lt;br&gt;Timing: Prenatal, Discharge to 8 wk, &gt;8 wk to 12 mo &lt;br&gt;Diagnostic comparator: None</td>
<td>Receipt of appropriate diagnostic/ treatment services for depression (detection rate, treatment rate)</td>
<td>Scores on diagnostic instruments for depression (referral rate)</td>
</tr>
<tr>
<td><strong>Zlotnick, 2006</strong>&lt;sup&gt;KQ 4&lt;/sup&gt;</td>
<td>Design: RCT &lt;br&gt;Location: U.S. &lt;br&gt;Setting: Prenatal care, short-term postpartum followup &lt;br&gt;Funding: Government Provider: NR</td>
<td>Assessed: 512 &lt;br&gt;Eligible: 201 &lt;br&gt;Enrolled: 99 &lt;br&gt;Completed: 86 &lt;br&gt;Analysed: 86</td>
<td>Sex: Female (99, 100%) &lt;br&gt;Mean age: 22.4 (SD 4.72) &lt;br&gt;Ethnicity: Hispanic or Latino N=44, Not Hispanic or Latino N=55 &lt;br&gt;Race: Asian N=2, Black/ African American N=17, White N=28, Other N=8 &lt;br&gt;Special population: None</td>
<td>Screening tool(s): 17-item postpartum depression risk survey &lt;br&gt;Timing: Prenatal &lt;br&gt;Diagnostic comparator: Longitudinal Interval Follow-Up Evaluation (depression module)</td>
<td>Scores on diagnostic instruments for depression</td>
<td>Maternal well-being/ parenting scores (Range of Impaired Functioning)</td>
</tr>
</tbody>
</table>

Abbreviations: ANRQ=Antenatal Risk Questionnaire; BDI=Beck Depression Inventory; BDI-II=Beck Depression Inventory-II; CIDI=Composite International Diagnostic Interview; CIDI-SF=Composite International Diagnostic Interview-Short Form; CIS-R=Clinical Interview Schedule, Revised; DIGS=Diagnostic Interview for Genetic Studies; DSM-III-R=Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition, Revised; DSM-IV=Diagnostic and Statistical Manual of Mental Disorders, 4th Edition; DSM-IV-TR=Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision; EPDS=Edinburgh Postnatal Depression Scale; GHQ-12=12-Item General Health
References to Appendix E


