Evidence-based Practice Center Systematic Review Protocol

Project Title: Core Needle and Open Surgical Biopsy for Diagnosis of Breast Lesions—An Update to the 2009 Report

I. Background and Objectives for the Systematic Review

Among women in the United States, breast cancer is the second most common malignancy (after skin cancer), and the second most common cause of cancer death (after lung cancer).[1] Approximately 1 in 8 U.S. women will develop breast cancer during their lifetime, and an estimated 2.7 million women had a current or past diagnosis of breast cancer as of 2009.[2] The American Cancer Society estimates that 232,340 new cases of invasive breast cancer and 64,640 new cases of noninvasive breast cancer will be diagnosed in 2013, and 39,620 women will die of breast cancer.[3]

During the earliest stages of breast cancer, there are usually no symptoms. The process of breast cancer diagnosis is initiated by detection of an abnormality through self-examination, physical examination by a clinician, or screening mammography. Data from the Behavioral Risk Factor Surveillance System show that, in 2010, 75.4 percent of U.S. women aged ≥40 years and 79.7% of women aged 50 to 74 years reported having a mammogram within the past 2 years.[4] If initial assessment suggests that the abnormality may be breast cancer, the woman may be referred for a biopsy, which is a sampling of cells or tissue from the suspicious lesion. Among women screened annually for 10 years, approximately 50 percent will need additional imaging tests, and a large proportion will have biopsies.[5, 6] More than a million women have breast biopsies each year in the United States. There are currently three techniques for obtaining samples from suspicious breast lesions: fine-needle aspiration, biopsy with a hollow core needle, or open surgical excision of tissue. Fine-needle aspiration, which retrieves a sample of cells, is generally considered less sensitive than both core-needle and open biopsy methods.[7] Core-needle biopsy, which retrieves a sample of tissue, and open surgical procedures are, therefore, the most frequently used procedures.

Samples obtained by any of these methods are evaluated by pathologists and classified into histological categories with the primary goal of determining whether the lesion is benign or malignant. Because a core-needle biopsy often samples only part of the breast abnormality, there is the risk that a lesion will be classified as benign or as high risk (e.g., atypical ductal hyperplasia [ADH]) or noninvasive (e.g., ductal carcinoma in situ [DCIS]) when invasive cancer is in fact present in unsampled areas. In contrast, open surgical biopsy often samples most or the entire lesion, and it is thought that there is a much smaller risk of misdiagnosis. However, while open surgical biopsy methods are considered to be the most accurate, they also appear to carry a higher risk of complications, such as bleeding or infection, when compared with core-needle biopsy.[8] Therefore, if core-needle biopsy is also highly accurate, women and their clinicians may prefer some type of core-needle biopsy to open surgical biopsy.

Source: www.effectivehealthcare.ahrq.gov
Published online: September 20, 2013

Core-needle biopsy may be carried out using a range of techniques. If the breast lesion to be biopsied is not palpable, an imaging method (i.e., stereotactic mammography, ultrasound, or magnetic resonance imaging [MRI]) may be used to locate the lesion. The biopsy may be carried out with needles of varying diameters, and one or more samples of tissue may be taken. Sometimes a vacuum device is used to assist in removing the tissue sample through the needle. It is thought that these and other variations in how a core-needle biopsy is performed may affect the accuracy and rate of complications of the biopsy. However, the impact aspects of biopsy technique have on test performance and safety are not clear.

In 2009, the ECRI Evidence-based Practice Center (EPC) conducted a comparative effectiveness review for core-needle versus open surgical biopsy on behalf of the Agency for Health Care Research and Quality (AHRQ).[9, 10] The review assessed the diagnostic test performance and harms of multiple core-needle biopsy techniques and tools, when compared with open surgical biopsy, and also evaluated differences between open biopsy and core-needle biopsy with regard to patient preference, costs, availability, and other factors. The conclusions were that core-needle biopsies were almost as accurate as open surgical biopsies, had a lower risk of severe complications, and were associated with fewer subsequent surgical procedures.[10] The need for an update of the 2009 review was assessed in 2010 by the RAND EPC.[11] Several high-impact general medical and specialty journals were searched, a panel of experts in the field was consulted, and an overall assessment of the need to update the review was produced. The conclusion of the updated Surveillance Report was that additional studies and changes in practice render some conclusions of the original report possibly out of date. Specifically, the Surveillance Report noted the following:

- New studies are available that could be included in the updated report regarding the following topics:
    - The underestimation rate of stereotactically-guided vacuum-assisted core-needle biopsy for DCIS
    - The test performance of MRI-guided core-needle biopsy
    - The test performance of freehand automated-device core-needle technology
- New studies on the test performance of core-needle biopsy may include additional information allowing the exploration of the heterogeneity for test performance or harm outcomes.

On the basis of the Surveillance Report findings, an updated review of the published literature was considered necessary to synthesize all evidence on currently available methods for core-needle and open surgical breast biopsy.

II. The Key Questions

The Key Questions (KQs) and study selection criteria (**p**opulation, **i**ntervention, **c**omparator, **o**utcome, **t**iming, and **s**etting; PICOTS) for this update began with those specified in the original report. On the basis of input from clinical experts during the development of this protocol, we have made selected revisions to the KQs and study eligibility criteria to clarify the focus of the updated systematic review.

The following three KQs will be addressed in the review:

Question 1

In women with a palpable or nonpalpable breast abnormality, what is the test performance of different types of core-needle breast biopsy when compared with open biopsy for diagnosis?

a. What factors associated with the patient and her breast abnormality influence the test performance of different types of core-needle breast biopsy when compared with open biopsy for diagnosis of a breast abnormality?
b. What factors associated with the procedure itself influence the test performance of different types of core-needle breast biopsy when compared with open biopsy for diagnosis of a breast abnormality?
c. What clinician and facility factors influence the test performance of core-needle breast biopsy when compared with open biopsy for diagnosis of a breast abnormality?

Question 2

In women with a palpable or nonpalpable breast abnormality, what are the harms associated with different types of core-needle breast biopsy when compared with open biopsy for diagnosis?

a. What factors associated with the patient and her breast abnormality influence the harms of core-needle breast biopsy when compared with the open biopsy technique in the diagnosis of a breast abnormality?
b. What factors associated with the procedure itself influence the harms of core-needle breast biopsy when compared with the open biopsy technique in the diagnosis of a breast abnormality?
c. What clinician and facility factors influence the harms of core-needle breast biopsy when compared with the open biopsy technique in the diagnosis of a breast abnormality?

Question 3

How do open biopsy and various core-needle techniques differ in terms of patient preference, availability, costs, availability of qualified pathologist interpretations, and other factors that may influence choice of a particular technique?

Study Eligibility Criteria

- **Population**

The population for all KQs is women who have been referred for biopsy for the diagnosis of primary breast cancer (including multifocal and bilateral disease) following self-examination, physical examination, or screening mammography. Studies carried out in women at high baseline risk of breast cancer (e.g., due to BRCA mutations) will therefore be included;

however studies carried out in women who have been previously diagnosed with breast cancer and are being examined for recurrence will be excluded.[a]

- **Interventions**

  For all KQs, the intervention is a core-needle biopsy done to evaluate whether a breast lesion is malignant. Other uses of biopsy techniques (e.g., use of biopsy to examine the sentinel lymph nodes in women with an established diagnosis of breast cancer) are excluded.

- **Comparators (reference standard and comparator index tests)**

  For test performance outcomes (KQ 1) the reference standard is either open surgical biopsy or followup by clinical examination and/or mammography for at least 6 months. The diagnostic performance of each core biopsy technique (each index test) will be quantified versus the reference standard.[b] The comparative diagnostic performance of alternative core-needle biopsy techniques is also of interest.[c]

  For harms and patient-relevant outcomes (outcomes other than diagnostic performance; KQs 2 and 3) the comparators are:

  - Open surgical biopsy,
  - Followup by clinical examination and/or mammography for at least 6 months
  - Alternative core-needle biopsy methods (e.g., stereotactic mammography vs. ultrasound to locate the breast lesion; use vs. nonuse of vacuum assistance to extract tissue samples)

- **Outcomes**

  - For KQ 1, test performance outcomes, as assessed by the following measures:

    - Sensitivity (proportion of cancerous tumors detected by the reference standard that are also detected by core-needle biopsy)
    - False-negative rate (proportion of negative findings according to core-needle biopsy that are classified as positive by the reference standard)
    - The underestimation rate for atypical ductal hyperplasia (ADH; proportion of core needle biopsy findings of ADH that are found to be malignant according to the reference standard)
    - The underestimation rate for DCIS (proportion of core-needle biopsy findings of DCIS that are found to be invasive according to the reference standard)

[a] The original review excluded studies carried out in women at high risk of breast cancer; however, magnetic resonance imaging (MRI)-guided biopsy, which has been identified as a topic of interest for the updated review, is used mainly in this subset of patients. For this reason, following extensive discussions with the TEP, we decided to broaden the scope of the review to cover women at high risk for cancer. In effect, this will be a de novo review with respect to this population subset.

[b] Most assessments of diagnostic performance quantify the sensitivity and the specificity of each index test—here each core-needle biopsy technique. Sensitivity and specificity are probabilities conditional on true disease status and are noncomparative in nature. The reference standard is used in their definition and is not a "comparator test."

[c] That is, differences or ratios of sensitivities and of specificities between alternative core-needle biopsy techniques.

- For KQ 2:

  - Rate of inconclusive biopsy findings (e.g., inadequate sampling of the lesion)
  - Repeat biopsy rate
  - Subsequent false-positive and false-negative rates on mammography
  - Dissemination (seeding) of cancerous cells along the needle track
  - Patient-centered outcomes (including bruising, bleeding or hematomas, pain, use of pain medication, infections, fainting or near fainting, and time to recover)

- For KQ 3:

  - Patient-relevant outcomes
    - Patient preferences for specific procedures
    - Cosmetic results
    - Quality of life
    - Anxiety and other psychological outcomes
    - Time to complete tumor removal (for women with cancer)
    - Recurrence rate (for women with cancer, including local, regional, and distant recurrence)
    - Cancer-free survival and overall survival

  - Resource use and logistics
    - Costs
    - Resource utilization other than cost (number of additional surgical procedures [e.g., re-excisions], procedural time)
    - Subsequent surgical procedures
    - Wait time for test results

  - Availability of technology and relevant expertise
    - Physician experience
    - Availability of equipment
    - Availability of (qualified) pathologists to evaluate biopsy samples

- **Timing**

  Duration of clinical and/or mammographic followup must be at least 6 months in studies where open surgical biopsy was not performed.

- **Setting**

  Studies in all geographic locations and care settings will be evaluated, including general hospitals, academic medical centers, and ambulatory surgical centers, among others.

III. Analytic Framework

The analytic framework is adapted from that published in Figure 1 of the original 2009 comparative effectiveness review.

Figure 1. Analytical framework

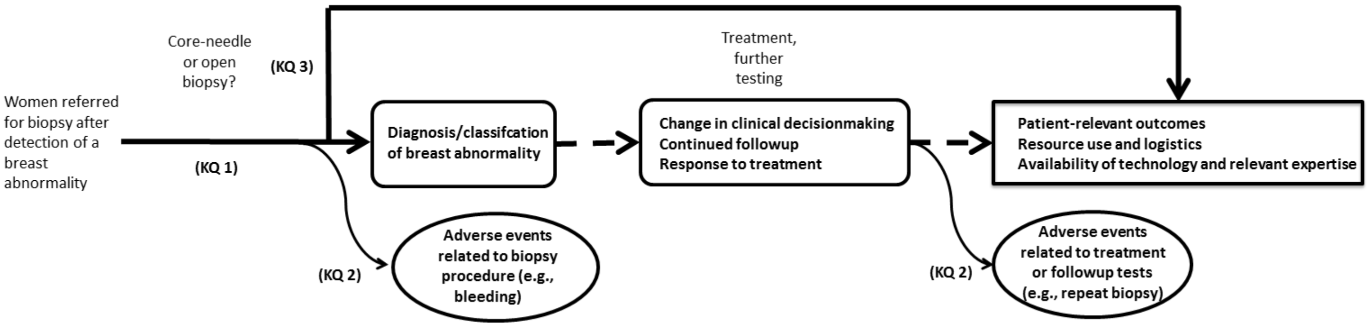
Women referred for biopsy after detection of a breast abnormality
Core-needle or open biopsy?
(KQ 3)
(KQ 1)
Diagnosis/classifcation of breast abnormality
Treatment, further testing
Change in clinical decisionmaking
Continued followup
Response to treatment
Patient-relevant outcomes
Resource use and logistics
Availability of technology and relevant expertise
(KQ 2)
Adverse events related to biopsy procedure (e.g., bleeding)
(KQ 2)
Adverse events related to treatment or followup tests (e.g., repeat biopsy)

Figure 1. This figure depicts the Key Questions (KQs) within the context of the PICOTS (**p**opulation, **i**ntervention, **c**omparator, **o**utcome, **t**iming, and **s**etting) described in the previous section. In general, the figure illustrates how core-needle biopsy versus open surgical biopsy may result in KQ 1 outcomes such as diagnosis of breast abnormalities, KQ 2 outcomes such as adverse events, and KQ 3 outcomes such as patient-relevant outcomes.

IV. Methods

A. Detailed Selection Criteria for Inclusion/Exclusion of Studies in the Review

The following inclusion/exclusion criteria apply to studies addressing **KQ 1** and are equivalent to the inclusion/exclusion criteria in the original report, unless otherwise noted:

1. Studies must have been carried out in women who have been referred for biopsy to detect possible breast cancer after an abnormality is found by self-examination, physical examination, or screening mammography. Studies in women at high risk of breast cancer were excluded from the original report but will be included in this update.
2. Studies must have used as reference standard for core-needle biopsy one or more of the following: open surgical biopsy and followup with clinical examination or mammography for at least 6 months.
3. Studies must have used biopsy instrumentation that is currently commercially available, as studies about discontinued devices are not applicable to current practice. Technologies excluded on this basis from the original review were the ABBI® (advanced breast biopsy instrumentation)

device, the MIBB® (minimally invasive breast biopsy) device, and SiteSelect® (stereotactic breast biopsy system). We will classify core-needle biopsy technology and equipment as outdated versus current through discussions with experts on the Technical Expert Panel (TEP), confirming that decisions to exclude certain technologies in the original report remain appropriate and checking the relevance of the technologies in more recent studies.

4. Studies must have been prospective or retrospective cohort studies or randomized controlled trials. Retrospective case studies ("case series"[12]) and other studies sampling patients on the basis of outcomes (e.g., diagnostic case-control studies, or studies selecting cases on the basis of specific histological findings) will be excluded. Empirical evidence from meta-epidemiological studies suggests that diagnostic case-control studies may overestimate test performance.[13]
5. Studies must have enrolled at least 10 participants per arm or per comparison group. This inclusion criterion is intended to reduce the risk of bias from nonrepresentative participants in small studies. Further, small studies do not produce precise estimates of test performance.
6. Studies must have followed at least 50 percent of participants to completion. This inclusion criterion is intended to reduce the risk of bias from high rates of attrition.
7. Studies must have been published in English. This inclusion criterion, which was also used in the original review, is due to feasibility constraints that do not permit translation of studies. There is no empirical evidence that excluding non–English-language studies from systematic reviews systematically biases review results.[14]
8. Studies must have been published in peer-reviewed journals as full articles. Meeting abstracts are not peer reviewed, typically do not provide complete information on study outcomes, and provide insufficient information to allow valid assessment of risk of bias.
9. Studies must include information on the sensitivity, specificity, positive or negative predictive values, or include data that allow the calculation of one or more of these outcomes. Specifically, studies need to provide adequate information to reconstruct 2 × 2 tables of test performance.

KQ 2 will be addressed by extracting harm-related information for core-needle biopsy and open surgical biopsy from studies addressing KQ 1. In addition, we will include studies that meet all other selection criteria for KQ 1, except for the use of a reference standard and the reporting of information on test performance outcomes. Specifically for KQ 2, we will use criteria 1 and 3–8, as listed above but will disregard criteria 2 and 9. This will allow us to consider additional sources of evidence that assess harms. Finally, we will collect any primary research articles, regardless of design (i.e., case reports and case series, case-control studies, cohort studies, and randomized trials), that address the dissemination of cancer cells by the biopsy procedure, a relatively rare harm that is specific to core biopsy.

The original report did not use formal criteria for study selection for **KQ 3.**[11] Based on the findings of the original report, we plan to use the same PICOTS criteria described above and consider the following study designs for this KQ:

- Cost studies, including cost-minimization and cost-consequence analyses, will be used to obtain information on resource utilization and unit costs. Given the large variability of cost information among different jurisdictions, we will only consider studies conducted in the United States.[15]
- Cost-effectiveness/cost-utility analyses based on primary trials[16] of breast biopsy interventions will be used to obtain information on unit costs and resource utilization. Specifically, we will consider the components of cost and resource use but will not use cost-effectiveness ratios or

other summary measures of cost-effectiveness/utility. As for cost studies, we will only consider cost-effectiveness/cost-utility studies conducted in the United States.[15] We will not use model-based cost-effectiveness results.

- Randomized controlled trials, cohort studies, and cross-sectional studies on patient preferences, cosmetic results of biopsy procedures, physician experience (including studies of the "learning curve" for different biopsy methods and tools) will be included.
- Studies of pathologist qualifications for interpreting core-needle biopsy results, including interlaboratory initiatives to standardize diagnostic criteria (e.g., proficiency testing) or minimal competency requirements, will be included.
- Surveys of the availability of equipment for obtaining core-needle biopsies and of qualified pathologists to examine biopsy samples will be included.

B. Searching for the Evidence: Literature Search Strategies for Identification of Relevant Studies To Answer the Key Questions

Appendix 1 describes the search strategy employed in the original review. This search will be adapted for use in MEDLINE®, EMBASE®, the Cochrane Central Register of Controlled Trials (CENTRAL), the Cochrane Database of Systematic Reviews, the Database of Abstracts of Reviews of Effects (DARE), the Health Technology Assessment Database (HTA), the National Health Service Economic Evaluation Database (NHS EED; United Kingdom), the National Guideline Clearinghouse™ (United States), and the Cumulative Index to Nursing and Allied Health Literature (CINAHL®) database.[17] We note that the original review used a filter for studies of diagnostic tests to increase the specificity of the search strategy; this is a reasonable approach, given the large volume of literature on studies on diagnostic biopsy methods for breast cancer. Because the update will have to cover a much shorter time period (from 2009 to 2013), we may opt to not use this filter to increase sensitivity.[18] Our searches will begin 6 months before the most recent search date in the original review to ensure adequate overlap. To identify studies excluded from the original review because they enrolled women at high risk for cancer, the set of abstracts screened for the original review will be obtained and rescreened for potentially eligible studies of high-risk women. In addition, the list of studies excluded from the original review following full-text review will be checked to identify studies excluded because they included women at high risk for breast cancer. We will also perform a search for systematic reviews on the topic and use their reference lists of included studies to validate our search strategy and to make sure we can identify all relevant studies.

All reviewers will screen a common set of 200 abstracts (in 2 pilot rounds, each with 100 abstracts) and will discuss discrepancies in order to standardize screening practices and ensure understanding of screening criteria. The remaining citations will be split into nonoverlapping sets, each screened by two reviewers independently. Discrepancies will be resolved by consensus with a third investigator.

Potentially eligible citations will be obtained in full text and reviewed for eligibility on the basis of the predefined inclusion criteria. Two reviewers will screen each article in full-text for eligibility. Disagreements regarding article eligibility will be resolved by consensus with a third reviewer. We will generate a list of reasons for exclusion for all studies excluded at the full-text screening stage.

We will ask the TEP to provide citations of potentially relevant articles. Additional studies will be identified through the perusal of reference lists of eligible studies, published clinical practice guidelines, relevant narrative and systematic reviews, Scientific Information Packages from manufacturers, and a search of U.S. Food and Drug Administration databases. All articles identified through these sources will be screened for eligibility against the same criteria used for articles identified through literature searches. If necessary, we will revise the search strategy so that it can better identify articles similar to those missed by our current search strategy. We will also ask the TEP to review the final list of included studies to ensure that no key publications have been missed.

Following submission of the draft report, an updated literature search (using the same search strategy) will be conducted. Abstract and full-text screening will be performed as described above. Any additional studies that meet the eligibility criteria, including those that are identified through the peer review and comment processes, will be added to the final report.

C. Data Abstraction and Data Management

Data will be extracted into standard forms. The basic elements and design of these forms will be similar to those we have used for other reviews of diagnostic tests and will include elements that address population characteristics, sample size, study design, descriptions of the index and reference standard tests of interest, analytic details, and outcome data. Prior to extraction, forms will be customized to capture all elements relevant to the KQs. We will use separate sections in the extraction forms for KQs related to short-term outcomes such as classification of breast abnormalities, intermediate outcomes such as clear surgical margins, patient-relevant outcomes such as quality of life, and for factors affecting (modifying) test performance. We will pilot test the forms on several studies extracted by all team members to ensure consistency in operational definitions. If necessary, forms will be revised before full data extraction.

Data from each eligible study will be extracted by a single reviewer. The extracted data will be reviewed and confirmed by at least one more team member (data verification). Disagreements will be resolved by consensus with a third reviewer.

While preparing the final evidence report, we will contact authors (1) to clarify information reported in the papers that is hard to interpret (e.g., inconsistencies between tables and text) and (2) to verify suspected overlap between study populations in publications from the same group of investigators. Author contact will be by email (to the corresponding author of each study), with a primary contact attempt (once all eligible studies have been identified) and up to two reminder emails (approximately 2 and 4 weeks after the first attempt).

D. Assessment of Methodological Risk of Bias of Individual Studies

We will assess the risk of bias for each individual study using the assessment methods detailed by the Agency for Healthcare Research and Quality in its *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*,[19] hereafter referred to as the *Methods Guide*. We will use the updated QUADAS 2 instrument to assess the risk of bias (methodological quality or internal validity) of the diagnostic test studies included in the review (these studies will comprise the majority of the available studies).[20-23] The tool assesses four domains for risk of bias related to patient selection, index test, reference standard test, and patient flow and timing. For studies of other designs, we will use appropriate sets of items to assess risk of bias or methodological "quality": for nonrandomized cohort studies we will use the Newcastle-Ottawa scale,[24] for randomized controlled trials we will use the Cochrane Risk of Bias tool,[25] and for studies of resource utilization and costs we will use the "Drummond checklist."[26]

We will not calculate "composite" quality scores. Instead, we will assess and report each methodological quality item (as Yes, No, or Unclear/Not Reported) for each eligible study. We will rate each study as being of low, intermediate, or high risk of bias on the basis of adherence to accepted methodological principles. Generally, studies with low risk of bias have the following features: lowest likelihood of confounding due to comparison to a randomized controlled group; a clear description of the population, setting, interventions, and comparison groups; appropriate measurement of outcomes; appropriate statistical and analytical methods and reporting; no reporting inconsistencies; clear reporting of dropouts and a dropout rate less than 20 percent; and no apparent bias. Studies with moderate risk of bias are susceptible to some bias but not sufficiently to invalidate results. They do not meet all the criteria for low risk of bias owing to some deficiencies, but none are likely to introduce major bias. Studies with moderate risk of bias may not be randomized or may be missing information, making it difficult to assess limitations and potential problems. Studies with high risk of bias are those with indications of bias that may invalidate the reported findings (e.g., observational studies not adjusting for any confounders, studies using historical controls, or studies with very high dropout rates). These studies have serious errors in design, analysis, or reporting and contain discrepancies in reporting or have large amounts of missing information. We discuss the handling of high risk of bias studies in sections E and F below.

In quantitative analyses, we will consider performing subgroup analyses to assess the impact of each quality item on the meta-analytic results. The grading will be outcome specific, such that a given study that reports its primary outcome well but did an incomplete analysis of a secondary outcome would be graded as having different quality for the two outcomes. Studies of different designs will be graded within the context of their study design. Thus, randomized controlled trials will be graded as having a high, medium, or low risk of bias, and observational studies will be separately graded as having a high, medium, or low risk of bias.

E. Data Synthesis

We will summarize included studies qualitatively and present important features of the study populations, designs, tests used, outcomes, and results in summary tables. Population characteristics of interest include age, race/ethnicity, density of breast tissue, and palpability of the lesion. Design characteristics include methods of population selection and sampling and followup duration. Test characteristics include imaging-guided versus not imaging-guided procedures and vacuum-assisted versus not vacuum-assisted biopsy methods. We will present information on test performance, harms, complications, patient preferences, and resource utilization including costs.

For KQs 1 and 2, we will judge for each outcome of interest whether the eligible studies are sufficiently similar to be combined in a meta-analysis on the basis of clinical heterogeneity of patient populations and testing strategies, as well as methodological heterogeneity of study designs and outcomes reported. For KQ 3, we will not perform a meta-analysis but will depict results graphically or in tabular form.

Based on discussions with local clinical experts and the original evidence report, we expect that eligible studies will have employed a variety of different biopsy methods (e.g., stereotactic guidance with vacuum assistance, ultrasound guidance with an automated system, MRI guidance). We will seek input from TEP members to define groups of "sufficiently similar" tests for synthesis (including meta-analysis) during later stages of the review. Of note, the material used to solicit TEP input will not include any data on outcome results extracted from the studies (to limit the potential for bias). The appropriateness of a meta-analysis will be determined before any data analysis; we will not base the decision to perform a meta-analysis on the statistical criteria for heterogeneity. Such criteria are often inadequate (e.g., low power when the number of studies is small) and do not account for the ability to explore and explain heterogeneity by examining study-level characteristics. Main analyses will include all relevant studies.

Subgroup analyses according to patient, breast lesion, biopsy procedure, clinician, and facility characteristics will also be performed. The concordance of findings across subgroup analyses will be evaluated qualitatively (in all instances) and quantitatively (using meta-regression, when the data allow). We will consider the following potential modifiers of test performance or other outcomes in meta-regression analyses: patient and breast lesion factors (e.g., age, density of breast tissue, microcalcifications, and palpability of the lesions), biopsy procedure factors (e.g., needle size, imaging guidance, vacuum extraction, and number of samples), and clinician and facility factors (e.g., training of the operator, general hospital vs. dedicated cancer facility). We will also perform subgroup analyses by individual risk-of-bias items to assess the impact of each on the results of the meta-analysis. We will evaluate the robustness of our findings in sensitivity analyses that exclude studies at high risk of bias. We will perform additional sensitivity analyses including leave-one-out meta-analysis, all-subsets meta-analysis, and a comparison of studies added in the update versus studies included in the original report.[27, 28]

Statistical analyses will be conducted using methods currently recommend for use in comparative effectiveness reviews of diagnostic tests.[29, 30] In cases where only a subset of the available studies can be quantitatively combined (e.g., when some studies are judged to be so clinically different from others as to be excluded from the meta-analysis), we will synthesize findings across all studies qualitatively by taking into account the magnitude and direction of effects.

F. Grading the Strength of Evidence for Individual Comparisons and Outcomes

We will follow the *Methods Guide*[19] to evaluate the strength of the body of evidence for each KQ with respect to the following domains: risk of bias, consistency, directness, precision, and reporting bias.[19, 31]

Briefly, we will define the risk of bias (low, medium, or high) on the basis of the study design and the methodological quality of the studies. Generally, the lack of studies at low risk of bias or inconsistencies among groups of studies at different levels of risk of bias will lead to downgrading the strength of the evidence. We will rate the consistency of the data as no inconsistency, inconsistency present, or not applicable (if there is only one study available). We do not plan to use rigid counts of studies as standards of evaluation (e.g., four of five studies agree, therefore the data are consistent); instead, we will assess the direction, magnitude, and statistical significance of all studies and make a determination. We will describe our logic in cases where studies are not unanimous. We will assess directness of the evidence ("direct" vs. "indirect") on the basis of the use of surrogate outcomes or the need for indirect comparisons. We will assess the precision of the evidence as precise or imprecise on the basis of the degree of certainty surrounding each effect estimate. This certainty is reflected in the confidence interval of each estimate. A precise estimate is one that allows for a clinically useful conclusion. An imprecise estimate is one for which the confidence interval is wide enough to include clinically distinct conclusions and that therefore precludes a conclusion.

We anticipate that most studies to be included in this update will be observational cohorts reporting on outcomes of test performance, using one or more index tests on all study participants. However, we also expect to find a small number of parallel-group, randomized or nonrandomized, comparative studies of alternative test strategies (e.g., reporting comparisons between alternative core-needle biopsy techniques). For patient populations (e.g., women at "average risk"), testing methods, and outcomes in cases where our work is a direct update of the previous systematic review, we will analyze all data jointly to reach conclusions that reflect the totality of the available evidence. For patient populations that were not covered by the original review (e.g., women at "high risk"), we will perform de novo syntheses. We will not combine the results of randomized and nonrandomized studies statistically; instead, we will qualitatively evaluate similarities and differences in study populations, diagnostic methods, and outcomes among study designs. We will use these comparisons to inform our judgments on the applicability of study findings to clinical practice (see also section G).

The potential for reporting bias ("suspected" vs. "not suspected") will be evaluated with respect to publication bias, selective outcome reporting bias, and selective analysis reporting bias. For reporting bias, we will make qualitative dispositions rather than perform formal statistical tests to evaluate differences in the effect sizes between more precise (larger) and less precise (smaller) studies. Although these tests are often referred to as tests for publication bias, reasons other than publication bias can lead to a statistically significant result, including "true" heterogeneity between smaller and larger studies, other biases, and chance, thereby rendering the interpretation of the tests nonspecific and the tests noninformative.[32, 33] Therefore, instead of relying on statistical tests, we will evaluate the reported results across studies qualitatively on the basis of completeness of reporting (separately for each outcome of interest), number of enrolled patients, and numbers of observed events. Judgment on the potential for selective outcome reporting bias will be based on reporting patterns for each outcome of interest across studies. We acknowledge that both types of reporting bias are difficult to reliably detect on the basis of data available in published research studies (i.e., without access to study protocols and detailed analysis plans). Because such assessments are inherently subjective, we will explicitly present all operational decisions and the rationale for our judgment on reporting bias in the draft report.

Finally, we will rate the body of evidence using four strength-of-evidence levels: high, moderate, low, and insufficient.[19] These will describe our level of confidence that the evidence reflects the true effect for the major comparisons of interest.

G. Assessing Applicability

We will follow the *Methods Guide*[19] to evaluate the applicability of included studies to patient populations of interest. We will evaluate studies (or subgroups of studies) of elderly women (operationally defined as patients 65 years of age or older) separately if data are available. Applicability to the population of interest will also be judged separately on the basis of patient characteristics (e.g., age may affect test performance because the consistency of the breast tissue changes over time), method by which suspicion is established (e.g., mammography vs. other methods may affect test performance through spectrum effects), baseline risk of cancer (women at "average risk" vs. those at "high risk" may affect estimated test performance because of differences in diagnostic algorithms), outcomes (e.g., prevalence of breast cancer diagnosed with a biopsy may also be a marker of spectrum effects), and setting of care (because differences in patient populations, diagnostic algorithms, and available technologies may affect test results).

V. References

1. American Cancer Society. Breast Cancer Facts & Figures 2011-2012. Atlanta, GA: American Cancer Society; 2011. Available at http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-030975.pdf.
2. National Cancer Institute Surveillance Epidemiology and End Results. SEER Stat Fact Sheets: Breast. Available at http://seer.cancer.gov/statfacts/html/breast.html. Accessed April 25, 2013.
3. American Cancer Society. Breast Cancer: What are the key statistics about breast cancer? Available at http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics. Accessed April 25, 2013.
4. Miller JW, King JB, Joseph DA, et al; Centers for Disease Control and Prevention. Breast cancer screening among adult women — Behavioral Risk Factor Surveillance System, United States, 2010. MMWR Morb Mortal Wkly Rep. 2012 Jun 15;61 Suppl:46-50. PMID: 22695463.
5. Hubbard RA, Kerlikowske K, Flowers CI, et al. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. Ann Intern Med. 2011 Oct 18;155(8):481-92. PMID: 22007042.
6. Elmore JG, Barton MB, Moceri VM, et al. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med. 1998 Apr 16;338(16):1089-96. PMID: 9545356.
7. Yu YH, Wei W, Liu JL. Diagnostic value of fine-needle aspiration biopsy for breast mass: a systematic review and meta-analysis. BMC Cancer. 2012;12:41. PMID: 22277164.
8. Agency for Healthcare Research and Quality. Having a breast biopsy: A guide for women and their families. AHRQ Publication No 10-EHC007-A. Rockville, MD: Agency for Healthcare Research and Quality; April 2010. Available at http://www.effectivehealthcare.ahrq.gov/ehc/products/17/407/core_needle_consumer_guide.pdf.
9. Bruening W, Schoelles K, Treadwell J, et al. Comparative Effectiveness of Core-Needle and Open Surgical Biopsy for the Diagnosis of Breast Lesions. Comparative Effectiveness Review No. 19 (Prepared by the ECRI Institute Evidence-based Practice Center under Contract No. 290-02-0019). AHRQ Publication No. 10-EHC007-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2009. Available at http://www.ncbi.nlm.nih.gov/books/NBK45220/pdf/TOC/pdf.
10. Bruening W, Fontanarosa J, Tipton K, et al. Systematic review: comparative effectiveness of core-needle and open surgical biopsy to diagnose breast lesions. Ann Intern Med. 2010 Feb 16;152(4):238-46. PMID: 20008742.
11. Maglione M, Motala A, Shanman R, et al; AHRQ Comparative Effectiveness Review Surveillance Program. Surveillance Report: Comparative Effectiveness of Core Needle Biopsy and Open Surgical Biopsy for Diagnosis of Breast Lesions. Rockville, MD: Agency for Healthcare Research and Quality; December 2011 and July 2012. Available at http://www.effectivehealthcare.ahrq.gov/ehc/products/17/370/NeedleBiopsy_SurveillanceAssessment_20120919.pdf.
12. Dekkers OM, Egger M, Altman DG, et al. Distinguishing case series from cohort studies. Ann Intern Med. 2012 Jan 3;156(1 Pt 1):37-40. PMID: 22213493.
13. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999 Sep 15;282(11):1061-6. PMID: 10493205.
14. Morrison A, Polisena J, Husereau D, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. Int J Technol Assess Health Care. 2012 Apr;28(2):138-44. PMID: 22559755.

15. Drummond M, Barbieri M, Cook J, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health. 2009 Jun;12(4):409-18. PMID: 19900249.
16. Ramsey S, Willke R, Briggs A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. Value Health. 2005 Sep-Oct;8(5):521-33. PMID: 16176491.
17. Whiting P, Westwood M, Burke M, et al. Systematic reviews of test accuracy should search a range of databases to identify primary studies. J Clin Epidemiol. 2008 Apr;61(4):357-64. PMID: 18313560.
18. Leeflang MM, Scholten RJ, Rutjes AW, et al. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol. 2006 Mar;59(3):234-40. PMID: 16488353.
19. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No 10(12)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; April 2012. Available at http://www.effectivehealthcare.ahrq.gov/ehc/products/60/318/MethodsGuide_Prepublication-Draft_20120523.pdf.
20. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36. PMID: 22007046.
21. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003 Nov 10;3:25. PMID: 14606960.
22. Whiting P, Rutjes AW, Dinnes J, et al. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004 Jun;8(25):iii, 1-234. PMID: 15193208.
23. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006 Mar 6;6:9. PMID: 16519814.
24. Wells G, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. [29 April 2013]; Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
25. Higgins JP, Altman DG, Gotzsche PC, et al; Cochrane Bias Methods Group/ Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011 Oct 18;343:d5928. PMID: 22008217.
26. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ. 1996 Aug 3;313(7052):275-83. PMID: 8704542.
27. Olkin I. Diagnostic statistical procedures in medical meta-analyses. Stat Med. 1999 Sep 15-30;18(17-18):2331-41. PMID: 10474143.
28. Olkin I, Dahabreh IJ, Trikalinos TA. GOSH–a graphical display of study heterogeneity. Res Synth Methods. 2012 Sep;3(3):214-23. DOI: 10.1002/jrsm.1053.
29. Trikalinos TA, Balion CM. Chapter 9: options for summarizing medical test performance in the absence of a "gold standard". J Gen Intern Med. 2012 Jun;27 Suppl 1:S67-75. PMID: 22648677.
30. Trikalinos TA, Balion CM, Coleman CI, et al. Chapter 8: meta-analysis of test performance when there is a "gold standard". J Gen Intern Med. 2012 Jun;27 Suppl 1:S56-66. PMID: 22648676.
31. Singh S, Chang SM, Matchar DB, et al. Grading a body of evidence on diagnostic tests. In: Chang SM, Matchar DB, Smetana GW, Umscheid CA, eds. Methods Guide for Medical Test Reviews.

AHRQ Publication No. 12-EHC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012:chapter 7. Available at http://www.ncbi.nlm.nih.gov/books/NBK98241/.

32. Lau J, Ioannidis JP, Terrin N, et al. The case of the misleading funnel plot. BMJ. 2006 Sep 16;333(7568):597-600. PMID: 16974018.
33. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002. PMID: 21784880.

VI. Definition of Terms

Selected terms related to this review and defined below are taken from the glossary published in the full report of the 2009 comparative effectiveness review.[9]

Atypical ductal hyperplasia. A condition in which the cells that line the milk ducts of the breast grow abnormally. The lesion itself is not malignant but may sometimes contain foci of malignant cells. Women with atypical ductal hyperplasia have an elevated risk of developing a malignant lesion.

Automated biopsy system. A device used to obtain core-needle samples. The device is pressed against the tissue at the appropriate location and angle and then the needle is “fired” into the tissue. After confirming the core-needle has sampled the appropriate tissue, the needle is withdrawn and the tissue sample ejected from the needle into a sampling container. Some units use a coaxial needle. With a coaxial needle, a cannula (hollow tube) is advanced into the tissue until reaches the area to be sampled; the sampling needle is then “fired” through the cannula and into the lesion.

Ductal carcinoma in situ. A carcinoma of the milk ducts of the breast that is confined within the duct.

Microcalcification. A tiny deposit of calcium visible as a bright spot on a mammogram. Tight clusters of microcalcifications may be a sign of a malignant lesion.

Stereotactic guidance. X-rays are taken from multiple locations in order to accurately identify the exact location of the lesion to be sampled. After using the images to determine where to sample, the needle is inserted. Further x-ray images are usually taken to confirm the needle has penetrated the lesion.

Ultrasound guidance. High-frequency sound waves are used to visualize the exact location of the lesion to be sampled. After using the images to determine where to sample, the needle is inserted. Images can be taken continuously during needle insertion to guide and confirm that the needle has penetrated the lesion.

Vacuum-assisted. After insertion of a hollow biopsy needle, a vacuum can be applied to pull tissue into the needle.

Underestimation rate. The percentage of lesions that were diagnosed with a core-needle biopsy as lesion types of lesser concern than the final diagnosis. For example, a lesion diagnosed as atypical ductal hyperplasia with a core-needle biopsy that is diagnosed as malignant with an open biopsy was "underestimated" by the core-needle biopsy.

VII. Summary of Protocol Amendments

There are no protocol amendments.

VIII. Review of Key Questions

For all Evidence-based Practice Center (EPC) reviews, the Key Questions were reviewed and refined as needed by the EPC with input from Key Informants and the Technical Expert Panel (TEP) to assure that the questions are specific and explicit about what information is being reviewed. In addition, the Key Questions will be posted for public comment and finalized by the EPC after review of the comments.

IX. Key Informants

Key Informants are the end-users of research, including patients and caregivers, practicing clinicians, relevant professional and consumer organizations, purchasers of health care, and others with experience in making health care decisions. Within the EPC program, the Key Informant role is to provide input into identifying the Key Questions for research that will inform health care decisions. The EPC solicits input from Key Informants when developing questions for systematic review or when identifying high-priority research gaps and needed new research. Key Informants are not involved in analyzing the evidence or writing the report and have not reviewed the report, except as given the opportunity to do so through the peer or public review mechanism.

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 are invited to serve as Key Informants and those who present with potential conflicts may be retained. The Task Order Officer (TOO) and the EPC work to balance, manage, or mitigate any potential conflicts of interest identified.

X. Technical Experts

Technical Experts comprise a multidisciplinary group of clinical, content, and methodological experts who provide input in defining populations, interventions, comparisons, or outcomes, as well as in identifying particular studies or databases to search. They are selected to provide broad expertise and perspectives specific to the topic under development. Divergent and conflicted opinions are common and perceived as healthy scientific discourse that results in a thoughtful, relevant systematic review. Therefore study questions, design, and/or methodological approaches do not necessarily represent the views of individual technical and content experts. Technical Experts provide information to the EPC to identify literature search strategies and recommend approaches to specific issues as requested by the EPC. Technical Experts do not do analysis of any kind nor contribute to the writing of the report and have not reviewed the report, except as given the opportunity to do so through the peer or public review mechanism.

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 are invited to serve as Technical Experts and those who present with potential conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential conflicts of interest identified.

XI. Peer Reviewers

Peer reviewers are invited to provide written comments on the draft report based on their clinical, content, or methodological expertise. Peer review comments on the preliminary draft of the report are considered by the EPC in preparation of the final draft of the report. Peer reviewers do not participate in writing or editing of the final report or other products. The synthesis of the scientific literature presented in the final report does not necessarily represent the views of individual reviewers. The dispositions of the peer review comments are documented and will, for Comparative Effectiveness Reviews and Technical Briefs, be published 3 months after the publication of the Evidence Report.

Potential Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Invited Peer Reviewers may not have any financial conflict of interest greater than $10,000. Peer reviewers who disclose potential business or professional conflicts of interest may submit comments on draft reports through the public comment mechanism.

XII. EPC Team Disclosures

The following team members will be involved:

- The EPC Director
- The EPC Codirector
- One Project Lead
- One Coproject Lead/Research Associate
- Two Research Associates
- One Project Manager
- One Program Assistant

All EPC team members have no financial or other conflicts of interest to disclose.

XIII. Role of the Funder

This project was funded under Contract No. HHSA 290-2012-0012-I from the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. The Task Order Officer reviewed contract deliverables for adherence to contract requirements and quality. The authors of the report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

Appendix 1: Search strategy for CINAHL/Embase/Medline from 2009 Comparative Effectiveness of Core Needle Biopsy and Open Surgical Biopsy for Diagnosis of Breast Lesions (Appendix B)

Set Number	Concept	Search statement
1	Breast biopsy	(breast biopsy or stereotactic breast biopsy or directional vacuum assisted biopsy).de.
2	Breast	Breast
3	Breast diseases	Exp breast cancer/di or exp breast neoplasms/di or exp breast disease/di or exp breast diseases/di
4		(breast or mammar$) and (Papilloma or calcification$ or calcinosis or tum?or$ or lesion$ or cancer or carcinoma$ or lump$)
5	Combine sets	or/2-4
6	Biopsy	5 and ((Biopsy or tumor biopsy).de. or biops$)
7	Large core needle biopsy	6 and ((needle biopsy or biopsy needle or percutaneous biopsy).de. or (large core or needle or mammotome or mammatome or vacuum))
8	Open biopsy	6 and (breast/su or breast tumor/su)
9		6 and (su.fs. or open or excision$ or incision$ or surgical)
10	Combine sets	8 or 9
11	Combine sets	or/1,7,10
12	Limit by publication type	11 not ((letter or editorial or news or comment or case reports or note or conference paper).de. or (letter or editorial or news or comment or case reports).pt.)
13	Diagnostics filter	12 and (exp prediction and forecasting/ or (predictive value of tests or receiver operating characteristic or ROC curve or sensitivity and specificity or accuracy or diagnostic accuracy or precision or likelihood).de. or ((false or true) adj (positive or negative)))
14	Clinical trials filter	13 and ((Randomized controlled trials or random allocation or double-blind method or single-blind method or placebos or cross-over studies or crossover procedure or double blind procedure or single blind procedure or placebos or latin square design or crossover design or double-blind studies or single-blind studies or triple-blind studies or random assignment or exp controlled study/ or exp clinical trial/ or exp comparative study/ or cohort analysis or follow-up studies.de. or intermethod comparison or parallel design or control group or prospective study or retrospective study or case control study or major clinical study).de. or Case control studies/ or Cohort/ or Longitudinal studies/ or Evaluation studies/ or Follow-up studies/ or Prospective studies/ or Retrospective studies/ or Case control study/ or Cohort analysis/ or Longitudinal study/ or Follow up/ or Cohort analysis/ or Followup studies/ or random$.hw. or random$.ti. or placebo$.mp. or ((singl$ or doubl$ or tripl$ or trebl$) and (dummy or blind or sham or mask)).mp. or latin square.mp. or (time adj series) or (case adj (study or studies) or ISRCTN$.mp. or ACTRN$.mp. or (NCT$ not nctc$)))
15	Combine sets	13 or 14
16	Eliminate overlap	
17	Seeding	12 and seeding.ti,ab.
18	Patient	12 and ((patient satisfaction or pain measurement or pain assessment or

	satisfactionQOL	visual analog scale or quality of life).de. or satisf$ or QOL or preference$)
19	Adverse events	12 and ((ae or co).fs. or cross infection or drainage or surgical wound infection).de.)
20	Disfiguration	12 and (disfigur$ or deform$)
21	Combine sets	Or/16-20