Comparative Effectiveness Research Review Disposition of Comments Report

Research Review Title: *Comparative Effectiveness of Core-Needle and Open Surgical Biopsy for the Diagnosis of Breast Lesions*

Research Review Citation: Bruening W, Schoelles K, Treadwell J, Launders J, Fontanarosa J, Tipton K. Comparative Effectiveness of Core-Needle and Open Surgical Biopsy for the Diagnosis of Breast Lesions. Comparative Effectiveness Review No. 19. (Prepared by ECRI Institute Evidence-based Practice Center under Contract No. 290-02-0019.) Rockville, MD: Agency for Healthcare Research and Quality. December 2009. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Comments to Research Review

The Effective Health Care Program encourages the public to participate in the development of its research projects. Each comparative effectiveness research review is posted to the Web site in draft form for public comment for a four-week period. Comments may be submitted through the Effective Health Care Web site or by postal mail. At the conclusion of the public comment period, comparative effectiveness research reviews are revised by the authors in response to the comments received.

Comments on draft reviews and the authors' responses to the comments will be posted publicly on the Effective Health Care Web site within three months after the final review is published. Comments are not edited for spelling, grammar, or other content errors. The table below includes the response by the authors of the review to each comment submitted for the draft review.

Section	Comment	Response
Executive Summary	Figures 2, 3, 4- perhaps including the formulas would be helpful?	Excellent suggestion. We have included formulas and definitions of the measures in the figure legends in the Executive Summary.
Executive Summary	Add a legend to figure 1 stating how many biopsies per method. Reviewer A60 also requested similar legends to be added.	We have made the suggested change.
Executive Summary	In the executive summary I think it would be important to indicate very briefly the ethnic diversity associated with breast cancer risk	While important, very few of the studies reported any information about the ethnic make-up of their enrolled patient populations.
Executive Summary	In executive summary second paragraph should read "mammography are x-rays of the breast using either film screen or digital images of the breast to masses, calcifications, or other abnormalities that could indicate invasive or non-invasive breast cancer	We have added information into the glossary to clarify how mammography may be performed. We decided to not make this change in the Executive Summary; the edits about how the x-rays are viewed are not of particular relevance to the report and unnecessarily complicate the sentence.
Executive Summary	Add architectural distortion to the BIRADS category 2	We did not make this change. Several other reviewers have pointed out that BI-RADS is copyrighted and the terminology should not be altered in any way
Executive Summary	Page 3 the word Ductal is missing	We have added the word.
Executive Summary	Under study selection you say the reference test is intended to measure the "true" disease status of each patient. Would it be more accurate to say the reference test is the one where the best evidence indicates it is the most accurate test available?	The statement about what the reference test is *intended* to do is then followed by explanations that available reference standards may not be able to achieve this goal and the implications.
Executive Summary	In the section on Pain the statement says 1.7% experienced severe pain; I'd suggest removing the "very" just state the facts	There is no "very" in the statement; perhaps the reviewer meant to remove "severe"? The studies measured pain on pain scales that included a category "severe" which some patients chose. Removing the word "severe" would render the statement untrue.
Executive summary	How are the authors defining the population of interest? What is the definition of a screening program? The patient population of interest is not clear.	We have edited the text in question for clarity.

Section	Comment	Response
Executive summary	The authors should define “routine physical examination” as a screening clinical breast examination. What about the diagnostic clinical breast examination done when women find an abnormality on breast self-examination? Please note that these findings are only relevant to about 50% of breast cancers if they do not include the patient-noted breast abnormalities.	The studies rarely reported data from these two populations separately; we have added text to clarify what the population of women is.
Executive summary	Patients referred after mammography should be analyzed separately from patients referred after detection of a palpable lump on physical exam.	The studies rarely reported data from these two populations separately; such an analysis is not possible. We did attempt to analyze data from palpable lesions separately from non-palpable lesions but were unable to come to any findings because of lack of reporting about lesion characteristics.
Executive summary	Executive summary needs a methods section and a discussion section	The template was revised to include Methods and Discussion sections.
Executive summary	The term “rate” implies a time factor. What is it?	We are aware that “rate” implies a time factor in epidemiological literature, but the literature on this topic very consistently uses the word “rate” to describe proportions in the context in which we used it. To change the terminology would be confusing to most readers. We have added additional definitions and formulas explaining the measures to the document.
Executive summary	P. 5 executive summary give assumed pretest probability.	It has been given in the footnote under the table, pre-test risk of 30%.
Executive summary	P. 4 executive summary please state how many more patients had severe bleeding events and how many fewer surgical procedures were performed.	The low quality of the evidence precludes an evidence-based estimate of the size of effect.
Executive summary	Please insert a footnote with a definition of “severe bleeding”.	The studies did not always provide a definition. Generally it was bleeding severe enough to require some type of treatment such as surgery or hospitalization. A footnote has been added to the table.

Section	Comment	Response
Executive summary	p. 3 executive summary also in results section please note how many patients were studied if possible.	It is not possible. Not all of the studies reported how many patients were enrolled; the number of breast lesions studied is as close as we can get.
Executive summary	Add United States to page 1 of the executive summary.	We have added the suggested text.
Executive summary	P. 1 Consider noting what USPSTF recommendations are in this paragraph.	Changes to the text suggested by other reviewers have taken care of this point.
Executive summary	p. 10 women with what kind of abnormalities on mammogram and/or PE?	We have added text to clarify the population of women who are being studied. Our inclusion criteria specified that we include "studies that enrolled women who were referred for biopsy after discovery of a possible breast abnormality by screening mammography or routine physical examination were included. Studies that enrolled subjects that were undergoing biopsy for any of the following purposes were excluded as being out of scope of the report: breast cancer staging, evaluation for a possible recurrence of breast cancer, monitoring response to treatment, evaluation of the axillary lymph nodes, evaluation of metastatic or suspected metastatic disease, or diagnosis of types of cancer other than primary breast cancer. Studies that enrolled patients from high-risk populations such as BRCA1/2 mutation carriers are also out of scope." In addition, the Background section includes the following: Breast cancer is usually first detected by feeling a lump on physical examination (either self-examination or an exam conducted by a health practitioner) or by observing an abnormality during x-ray screening mammography."
Executive summary	p. 15 indicate what the numbers in the figure mean	Figure legend added
Introduction	Add image guided core or stereotactic core biopsy to the list of methods used to collect the biopsy material	We did not make this change. The different types of core-needle biopsy are described later; the

Section	Comment	Response
		sentence in question talks about major categories of methods to collect biopsy material.
Introduction	Another important point to raise in this section is that a history of previous benign breast biopsy is a significant risk factor for subsequent breast cancer	In their 2007 *Cancer Epidemiology Biomarkers & Prevention* paper, Ashbeck et al. demonstrate that a history of previous benign breast biopsy is not a risk factor for subsequent cancer; they demonstrate that women diagnosed with certain types of benign lesions (e.g., ductal hyperplasia) have an elevated risk, whereas women diagnosed with other types of benign lesions (e.g., fibroadenomas) are at normal risk of developing subsequent breast cancer. We mention the various types of benign lesions that are associated with elevated risk of breast cancer in the Background section of the report.
Introduction	DCIS classifications use histologic grade, so the statement that only invasive tumors are described by histologic grade is inaccurate. Reviewer A59 also made this comment.	We have edited the statement.
Introduction	It is unclear why a specific manufacturer and device name are included in the background section. These should be deleted. Reviewer A59 made the same comment.	We had thought that the inclusion of familiar device names would help clarify the various types of biopsy. As this does not appear to be the case, we have deleted them.
Introduction	Page 42 mammographic screen should be changed to mammogram.	We have made the change.
Introduction	Page 42 insert more in front of expensive	We have made the change.
Introduction	Page 42 change general screening purposes to general mammography	We have made the change.
Introduction	BI-RADS® is a registered trademark and must always be written that way. Furthermore the terminology are copyrighted and should not be modified. Reviewer A59 made the same comment.	We have made the suggested modifications. Just as a side note, these suggestions do not appear to be commonly followed in the literature.
Introduction	p. 9 insert “usually” before “clinically”	We have made the suggested change.

Section	Comment	Response
Introduction	p. 9 change abnormal densities to “abnormalities”	We have made the suggested change.
Introduction	p.9 insert “close to” before 100%	We have made the suggested change.
Introduction	P. 9 change some to many, and breast cancer screening to annual mammography	We have made the suggested change.
Introduction	p. 23, correct spelling of vacuum	We have made the suggested change.
Introduction	p. 10 BIRADS 3 or higher is incorrect for screening mammography. Most abnormal examinations are coded as BIRADS 0. BIRADS 3 is only used after additional workup.	We agree that final assignment of a diagnostic BIRADS 3 category isn’t recommended until after further workup. However, often the screening mammogram is assigned a BIRADS number of 3 to 5 before further workup is performed.
Introduction	Delete MRI here as it is not recommended for the initial work-up off of screening mammography. We also don’t routinely do a physical examination.	This section was not intended to only refer to the initial workup. We have edited the text for clarity.
Introduction	p. 10 triple assessment usually only refers to clinical examination, imaging, and fine-needle aspiration	We concur and have altered the wording for clarity.
Introduction	p. 10 insert e.g. before stereotactic mammography	We have made the suggested change.
Introduction	p. 10 insert “biopsy” before procedure	We have made the suggested change.
Introduction	p. 10 add MRI	We have made the suggested change.
Introduction	p. 10 some vacuum devices require re-insertion	We have edited as suggested.
Introduction	p. 11 large core is vague	We have changed all references to non-large-core devices to core-needle throughout the document for clarity. Large-core devices use needles that are

Section	Comment	Response
		much larger (so large the diameter of the bore is usually measured in millimeters rather than gauge) than those used for core-needle biopsy (generally 11G or smaller).
Introduction	p. 11 some of these tumor types aren't "basic"	We have removed the word "basic".
Introduction	p. 11 MRI is often used for local staging of breast cancer; also add PET/CT as an option for staging	We have edited the sentence.
Introduction	p. 12 insert distant before metastasis	We have made the suggested change.
Introduction	p. 12 change suggests to shows and may be to usually	We are concerned that these wording changes imply greater certainty than can be determined at this time. We have acknowledged that for the purpose of analysis, it is necessary (and common practice) to make this assumption, but want to be clear that it is an assumption.
Introduction	p. 12 add "or lesion" after individual patient	We have made the suggested change.
Introduction	The USPSTF and many other organizations suggest starting screening at age 40 years of age, not age 50.	We are aware of this and have altered the text slightly for clarity. Some organizations question this recommendation; however, all organizations agree women age 50 and older should be screened. Thus the wording used in our report. Note: the updated USPSTF recommendations were inserted into the final version of the report after the review process was finished.
Introduction	It should be made clear from the beginning that fine needle aspiration is not the topic of the report.	We have added a sentence to clarify this point.
Introduction	The authors should clarify that women with BIRADS5 may have a recommendation from the radiologist to get a biopsy without first undergoing additional imaging.	We did not make the suggested change. Although true, we thought it would make the text confusing, and the point is not particularly relevant to the focus of the report.

Section	Comment	Response
Introduction	The process that breast biopsy works within needs to be better described so the reader understands the pros and cons better. The discussion needs to focus more on the clinical implications of the biopsy approaches.	We have added the additional text the reviewer provided.
Introduction	p. 10 BIRADS 0,4, 5, and sometimes 3	Text modifications in response to other reviewer suggestions has taken care of this point.
Introduction	p. 10 change “area may be” to “area is”	We did not make the suggested change because it reduced the overall accuracy of the statement.
Introduction	p. 43 the entire lesion, and samples lymph nodes, thus being	We made the suggested change.
Methods	An important weakness in this report is not addressing false-positive harms. This is not a minimal issue as it is associated with over-treatment harms. The paragraph under Data Analysis and Synthesis minimizes the harms associated with false-positive exams. Reviewer A60 also commented on this point. I think that not including harms of overdiagnosis is a problem. I think that women are undergoing unnecessary mastectomies and this should not be ignored.	We have modified the paragraph to indicate that there are harms associated with false-positive exams. However, as discussed in our Methods section, the literature on core-needle breast biopsy accuracy generally assumes that false-positives do not exist. Diagnoses of malignancy on core-needle biopsy that cannot be confirmed on open biopsy are assumed to have been fully removed by the core-needle biopsy. Whether this assumption is true or not cannot be directly addressed with the current literature base. We have a detailed discussion of the possible impact of the assumption that false-positives do not occur; and we believe that even if this assumption is incorrect, our conclusions about accuracy of diagnosis are not significantly incorrect. The possibility of cases of over-treatment occurring is out of the scope of the current report, which does not address treatment. It would be unusual for a woman to undergo a mastectomy following a positive core-needle biopsy; generally an open biopsy to confirm the core-needle biopsy result would be performed first. False-positive findings on core-needle biopsy may lead to a few unnecessary open biopsies being performed, but, in general, the use of core-needle biopsy reduces the overall number of open biopsies.

Section	Comment	Response
Methods	The assumption that pathologists are 100% accurate is a significant weakness. Reviewer A60 also commented on this point.	Pathology errors made while examining the core-needle specimens are captured in our analysis of core-needle biopsy accuracy- misdiagnoses may be due to poor sampling and may also be due to incorrect reading of the retrieved specimen. We are assuming that the open surgical biopsies are 100% accurate. We know this statement is not true, but it is very close to the truth. We discuss in detail the possible impact of this assumption on our conclusions about diagnostic accuracy. Even if this assumption is incorrect, our conclusions about accuracy of diagnosis are not significantly affected. We have added statements indicating that pathology errors are, however, not clinically insignificant due to the possibility of harmful over- and under-treatment.
Methods	Pathologists are often unable to distinguish between ADH and DCIS and assuming they are 100% correct is a significant weakness	In our primary analysis, a core-needle specimen that was diagnosed as ADH but was actually DCIS would be considered an error- an "ADH underestimation." Due to lack of reporting of details by the primary studies we were unable to perform any further analyses to identify core-needle accuracy in distinguishing between these two types of lesions. In the clinic, failure to distinguish between the two types of lesions on core-needle biopsy would be unlikely to have any impact on patients or medical practice because patients diagnosed as either ADH or DCIS are usually referred for open biopsy and surgical removal of the lesion.
Methods	I am not sure anyone would agree these days that open surgical biopsy is the "gold standard" of evaluating a suspicious breast lesion because the stereotactic cores are much less invasive and as accurate	The term "gold standard" is used to refer to a method of diagnosis that is as close to 100% accurate as possible. It is not being used to refer to the "gold standard of care". We have added the word "reference" and additional text to clarify.
Methods	Most researchers would use the term false negative rates rather than errors	We used the word "error" in an attempt to explain in plain language what the concept of sensitivity means.

Section	Comment	Response
Methods	I am not sure why they are using negative likelihood ratio rather than false negative rates	We are using both concepts. The concepts of false-negative rate are captured in the definition of sensitivity. Negative likelihood ratio is a totally different concept and measure. We have added additional information to the report to aid the reader in understanding the diagnostic terms used throughout the report.
Methods	Instead of negative likelihood ratio I think you mean false negative ratio	The correct terminology is negative likelihood ratio. As mentioned above, we have added a brief primer on diagnostic test characteristics to the Background section to aid reader understanding of the report.
Methods	You should justify the use of false negative ratio instead of false negative rate	We did not use either measure in our report, see response to previous two comments.
Methods	The underestimation rates need to be better defined.	We have changed all mention of underestimation rates to percentages to clarify the situation and added additional information, including a table, into the text to aid readers in understanding the outcome measures.
Methods	Your description of negative predictive value as applying to specific populations of women is not clear to me.	Negative predictive value varies as the prevalence of disease varies. We have made some edits to the text to help clarify the situation.
Methods	If you included lobular carcinoma and DCIS as in situ tumors you should clarify this in methods under assessment of performance	Unfortunately the degree of detail reported by many of the studies precludes us from, in many cases, being sure. Many of the studies simply reported "benign" "malignant" categories, including in situ types with invasive to get the malignant category; others reported "benign", "in situ" and "invasive" categories.
Methods	Many tumors have both DCIS and invasive features and it would be nice to know how often this occurs.	See above comment. Most such tumors would probably be classified as "invasive" in most of the studies.
Methods	The phrasing of the outcome "missed cancers" needs some clarification; does it include underestimates?	Not in the analysis of sensitivity. Underestimates would have been counted as true positives. Missed cancers were defined as women diagnosed as "benign" who were later found to not be so. We

Section	Comment	Response
		have added text and tables and formulas to clarify.
Methods	Have the authors considered showing both the mean and median quality score?	In our experience, quality scores are not distributed normally and the median score is a more accurate measure of the central tendency.
Methods	Consider adding figures or tables or other to clarify the definitions of underestimation rates and false positives, etc.	This is an excellent suggestion. We have added additional graphics to illustrate the point.
Methods	I am not sure I would consider a core-needle diagnosis of ADH as a true positive if cancer is found on the open.	These kinds of errors are captured in the ADH underestimation rate; most people consider this to be a "true positive" because no cancers were missed, because ADH diagnoses are generally referred for open biopsy.
Methods	"the chain of evidence linking…" I have no idea what this means.	We have extensively edited that statement to clarify its meaning.
Methods	Have the authors considered the possibility that with core biopsy becoming more common in the US, radiologists may be more likely to obtain core biopsies given the lower harms associated; and the more you look, the more likely you will find ADH, leading to more open procedures?	Over-treatment is always a concern. We agree that monitoring for this possibility would be worthwhile and hope that future research will clarify best practices for follow up to findings of ADH and other high-risk lesions on core biopsy.
Methods	p. 12. Explain what Level 2 means in the text.	We have added additional details.
Methods	Clarify what the reference standard is. Why did the authors choose a six month followup period?	We have added additional text to clarify. We are aware that 2 years is commonly suggested in the literature; however, requiring 2 years would have excluded the majority of the literature on a criterion that has not been proven to be necessary- it is expert opinion. We performed meta-regressions and found that studies that used open surgical biopsy on all patients, studies that used 2 years or longer followup on all patients, and studies that used 6 months or longer followup on all patients all reported very similar data on the accuracy of core-needle biopsy.

Section	Comment	Response
Methods	p. 21. What is meant by small lesions?	We have modified the text to clarify.
Methods	p. 18 add "for Question 3 due to the nature of the question and lack of published data on the topic"	We did not make this change. Inclusion criteria were developed *a priori* before searching the literature; at the time of inclusion criteria development the extent of or lack of published literature was unknown.
Methods	For funding- what abbreviation will the authors use if there is no external funding vs. none reported? This is an important distinction.	We did not encounter any publications that clearly stated there was no external funding. The question doesn't ask about external vs. internal funding; it asks if the study was funded by a source that clearly had a financial stake in the outcome of the study. If a study reported it was internally funded we would have answered "no" due to concerns about the center making money from performing the core-needle biopsies being studied and therefore being possibly biased towards finding the biopsies are accurate and safe. We also agree that lack of reporting of funding sources is a problem with quality assessment, as is the lack of a mechanism for dealing with less tangible benefits that may bias study outcomes, such as professional advancement. Additional methodology research may develop mechanisms to incorporate these concepts into quality assessment.
Methods	p. 17 change to biopsy for any of the following purposes	We have made the suggested change.
Methods	The grading of strength of evidence is unclear to me.	The description of how the strength of evidence was graded was accidently omitted from the draft report. This error has been corrected.
Methods	There are some issues with the methods section specifically re: the quality ratings.	The methods describing the grading of the quality of the evidence were accidently deleted during preparation of the draft. We have corrected the error.

Section	Comment	Response
Methods	Methods section not clear about when non-randomized studies searched for and where these studies have been applied.	The inclusion criteria indicate the study designs that were searched for and included. Randomization was not required. Randomized controlled trials rarely exist in the diagnostic literature and are often not as appropriate as diagnostic cohort studies in addressing questions of diagnostic accuracy.
Methods	What studies were excluded for KQ2 + 3?	KQ2 was answered using studies that met the inclusion criteria for KQ1, and for KQ3 we used any study that came up in the literature searches that reported relevant information. These points are mentioned in the Methods section.
Methods	It looks like non-trials were included for KQ2 +3 so please say something about how you searched for these.	The searches for literature addressing all questions are exactly as described. We did not include non-trials for KQ2 as is stated in the text. KQ3 was at times supplemented with expert opinion (e.g., the Technical Expert Panel and review articles identified by the described literature reviews) as stated in the text.
Methods	I assume that included studies also included women who had benign and malignant lesions- what is so different about these studies that they were excluded?	Some studies specifically selected patients to enroll on the basis of the final diagnosis of the lesion (retrospective case studies). Such studies are subject to a high probability of selection and spectrum bias and were excluded on that basis.
Methods	Replace Inconclusive with Insufficient	We have done so.
Methods	Add a reference to STATA and state what version was used	We have made the suggested changes.
Methods	A clear descriptive narrative of the application of what components of the AHRQ method's manual should be added.	The current draft of the AHRQ Methods Manual does not contain a chapter on Diagnostics. In its current form very little of the manual is applicable to this report.
Methods	A description on strength of evidence should be added.	The methods describing the grading of the quality of the evidence were accidently deleted during preparation of the draft. We have corrected the

Section	Comment	Response
		error.
Methods	Add the range of the quality score where it is first mentioned.	We have added it.
Methods	The reference standard is not 100% accurate and yet they used the term “sensitivity”.	No reference standard is 100% accurate. The reference standard used in this report is much more accurate than most. We discuss at length the impact of errors in the reference standard on our results. Most diagnostic methodologists feel it is usually incorrect to use correlations when analyzing diagnostic accuracy data because correlations cannot distinguish between the effects of false-positive and false-negative errors, nor can they account for the tradeoff between these errors, nor can they account for threshold effects. A number of methods have been invented to deal with the impact of very inaccurate reference standards on diagnostic analyses, see for example Phelps and Hutson 1995.
Methods	Keep the term “sensitivity” and only add in footnote “in comparison to reference test”	The definition of sensitivity is always “in comparison to a reference test”. You cannot calculate sensitivity without comparing the investigational test to a reference test. We have added a short primer on diagnostics to the Background section to aid readers in understanding the report.
Results	The section on availability of qualified pathologists is inadequate; I fear that the search terms were incorrect. Searches on pathologist diagnostic agreement may yield information.	Pathologist agreement and availability of qualified pathologists are not the same concept. This section is attempting to explore the issue of persons living in medically under-served areas who may not have access to experienced pathologists.
Results	You comment on the training and characteristics of the radiologists but not on the pathologists or whether double reading was used.	Very few of the studies reported any information about the pathologists or whether double reading was used.
Results	Please distinguish, in clinician characteristics, whether the data were insufficient for radiologists, surgeons, or pathologists.	The data was insufficient for all categories of clinicians.

Section	Comment	Response
Results	You note that the sensitivity of core-needle biopsy improved as the operators gained experience. More detail is needed.	We have added the requested additional information.
Results	I am shocked at the variability that exists for the number of missed cancers and this deserves comment in the conclusion.	The variability almost certainly comes from pooling data from studies that may have differed in many small details such as years of experience, exact procedures performed, and types of patients enrolled. The confidence intervals around some of the outcome measures are reasonably narrow. The measures that have wider confidence intervals are those derived from the smaller evidence bases.
Results	The article Poplack et al. 2005 should be added to the relative costs section	The article by Poplack et al. is certainly interesting but doesn't present relative costs of different procedures in a way that can be interpreted usefully for the current report.
Results	The scope in terms of clinician factors needs clearer delineations.	We included any clinician factors reported by the studies.
Results	The section Negative Surgical Excision after Needle Core Biopsy should have the words Core and Needle swapped	We have made the suggested change.
Results	In the section on dissemination of cancer cells during the biopsy procedure the risks of breast cancer that are associated with prior benign breast biopsies should be discussed.	We did not make this change; see our previous response to the comment about risk of breast cancer following a benign breast biopsy. The section in question is about dissemination of cancerous cells by biopsy procedures. Patients who have biopsies of benign lesions most likely do not have any cancerous cells to be disseminated.
Results	One issue not addressed was that PPVs of core biopsy performed by surgeons are much lower than those in the radiological literature.	We attempted to address this issue but were unable. The majority of the studies did not report whether surgeons or radiologists were performing the core biopsies.
Results	P. 19 there is some literature to support the concept that local pathologists have a very high agreement with central experts- Collins et al. 2004	Yes, we have cited Collins et al. 2004 in the Results section, Key Question 3

Section	Comment	Response
Results	p. 23, add discussion of lesion type from Fajardo et al. ref. 45 where calcifications were less accurately sampled than masses. This is currently discussed on page 25 under "multiple methods" but is easily overlooked there.	p.23 is data specifically about stereotactic automated gun biopsies; Fajardo et al. used several different methods of biopsy and did not report results separately by biopsy method.
Results	p. 23, Liberman et al. 1994 have a study on how many cores are enough this should be cited and discussed on lines 20-21.	Liberman et al. 1994 was excluded from consideration because they did not verify if any of their core-needle diagnoses were correct by any reference method.
Results	There are at least three studies of MRI-guided biopsy: Perlet et al. 2006; Lehman et al. 2005; and Lee et al. 2007	Perlet et al. 2006 and Lee et al. 2007 were excluded from consideration because they studied a population of women at very high risk of breast cancer- the results of their studies are not applicable to the general population. Lehman et al. 2005 was excluded from consideration because they did not verify their diagnoses of "benign" on core-needle biopsy.
Results	p. 26 add information from Liberman et al. 2001	Liberman et al. 2001 was excluded from consideration because they did not verify if their core-needle diagnoses were correct by any reference method.
Results	p. 43 add increased vasovagal reactions when discussing upright vs. prone units	We were unable to find any such discussion on or near p. 43. However, we have noted in the document that vasovagal reactions are more likely in upright units.
Results	p. 43 not sure what is meant by device.	We changed the word "device" to "system".
Results	p. 43 only two devices require the purchase of an expensive console. One manufacturer doesn't charge for the console if the user agrees to purchase a certain number of probes. Another is $10,000. There are also disposable, self-contained vacuum-assisted devices that have no console that cost $270 each. Hence the statement that the devices cost $37,000 is highly inaccurate.	We have modified the text to explain what exactly the $37,000 charge is referring to.
Results	Conclusion statement is confusing	We removed the confusing word core-needle from the statement.

Section	Comment	Response
Results	In the discussion move (24.4%, 27.1%) after our review and add the word “respectively”	We have made the suggested change.
Results	p. 23 The authors state that 29 of the studies were conducted at a single center. They need to state how many were conducted at more than one center.	We have added text to clarify this statement.
Results	p. 39 “three vacuum-assisted biopsy procedures” add the total number of procedures to this sentence	We have made the suggested change.
Results	How are the authors defining some of these harms in the original papers? What is “troublesome bleeding”? What is acute inflammation? Consider defining “satisfaction” and “unacceptable”. the authors should add some sort of definitions.	Unfortunately the original papers did not offer definitions in most cases. Where necessary we used the exact terminology, such as “troublesome bleeding”, used in the original papers.
Results	p.39 add more description of the methods used in these three studies. Were the radiologists told that there was a biopsy and did they know the specific location?	See various comments about the poor level of reporting in the original studies.
Results	p. 42 “within the last five years” is too vague. The authors should specify the years to aid readers using this report a decade later. Are they referring to a publication date or the date of the cost estimates?	The publication date. We have added specific dates to the text.
Results	p. 44 Collins et al. does not report that the agreement for ADH was near 100%.	Yes, but overall agreement for all lesions in Collins et al. was near 100%. We have added additional information about the ADH issue to the text. The disagreement over ADH diagnoses generally involved ADH/DCIS disagreements. ADH/DCIS disagreements are not likely to impact clinical practice or patients, because either diagnosis on a core-needle biopsy usually results in a referral for open surgical biopsy.
Results	p. 45 Did the authors consider the possibility that facilities that offer core biopsies might be larger cancer hospitals and thus have shorter wait times for results?	There was no evidence presented in the studies to support this hypothesis.
Results	P. 50 I found the summary hard to read visually. Consider putting into a table.	We have added a summary table.

Section	Comment	Response
Results	P. 50 You need to specific if you are referring to a BIRADS 4 on a screening mammogram or a diagnostic mammogram as they will have a different PPV. Do the authors mean the PPV or the pretest probability before the results of the test are known? The PPV is different for screening detected vs. diagnostic evaluations and there are three different definitions of PPV to distinguish between. Clearly state if you are using PPV1, PPV2, or PPV3 definition.	Nowhere do we refer to PPV of any definition. However, we believe the reviewer is actually asking for clarification of the study populations enrolled. We have added additional text to clarify the population under study.
Results	p. 50 please clarify that both core and open biopsies are done on 100% of the women and you will find DCIS/invasive in 1% or less of the open biopsies among the women with a normal core biopsy. I also do not understand how you obtained these data that are specific to BIRADS 4 from the publications reviewed.	We included women who had followup rather than open biopsy, as indicated in the inclusion criteria. We did not obtain any data "specific" to BIRADs-4 from the publications; we used the likelihood ratios calculated from the publications and information about BIRADS-4 reported in the general literature (citations are provided) to generate these numbers. We have added additional information to the Background section to aid the reader in understanding and using likelihood ratios.
Results	The wide range of prevalence of malignancy from 15% to 68.9% suggests to me that the 15% is among women referred after screening mammography for quick US guided biopsies, while the 68% are the women who have patient- or clinician- noted breast abnormalities.	This may be true; however, the reported data are insufficient to substantiate this speculation.
Results	The authors need to add the percent of patients not satisfied with cosmetic results, as this is a summary of harms.	The cosmetic result section is presented in the answer to Key Question 3 due to the explicit wording of the question, which asks for cosmetic results to be presented there.
Results	The study does not seem to separate out calcification findings from mass findings.	We attempted to do so but the majority of the studies did not report their data in a way that allowed us to come to any conclusions about different types of lesions.
Results	p. 50 the evidence seems to suggest a risk of malignancy of much less than 8%.	The 8% comes from the upper bound of the 95% confidence interval, which, since the confidence interval is fairly wide, is higher than the point estimate.

Section	Comment	Response
Results	The evidence seems strong that open surgical biopsies carry higher risks of complications.	We have reworded the conclusion; the Low rating was intended to only rate the estimate of size of effects for core-needle biopsy.
Results	I'm not sure of controversy. Current standard of care is that ADH be subject to open biopsy.	We encountered a number of publications that questioned whether the current standard of care for ADH was the best approach.
Results	The ACR has guidelines and technical standards on breast imaging and biopsies.	Yes; the current report, however, does not discuss any of the many clinical practice guidelines on breast imaging and biopsies.
Results	I think under patient characteristics I would add more about race/ethnicity	Very few of the included studies reported any details about race or ethnicity of their enrolled patients.
Results	I would include information about insurance coverage (e.g. difference in copayments).	This information was not included because of the wide variation and fluctuations in coverage policy and out-of-pocket costs for diagnostic procedures. Furthermore, a comprehensive search for coverage policies was outside the scope of this review.
Results	Add strength of evidence ratings to Key Question 3.	Due to the nature of KQ3, we did not address it in a formal evidence-based fashion. It was clear from the beginning of the project that there would be no specific studies available that were formally studying any aspect of KQ3. Therefore we approached the question as an "opinion/discussion" question rather than an "evidence" question. For this reason we did not rate the strength of the evidence nor did we draw any formal evidence-based conclusions.
Results	Change United Kingdom to UK in the table	We have made the suggested change.
Tables	No additional comments submitted.	
Figures	No additional comments submitted.	

Section	Comment	Response
Discussion	The sections on limitations of the evidence base and future research need more detail	We have added more details and significantly expanded these sections.
Discussion	The future research section should be enlarged.	We have added to the Future Research section.
General	Several typographical errors were pointed out	We have corrected the typographical errors.
General	I would like to have seen references in many locations throughout this report.	Without specific suggestions we are unable to respond to this comment; we have cited many literature sources throughout the report and are unaware of any statements that require additional references. Perhaps the reviewer is referring to the Executive Summary, where citations were deliberately omitted? That material plus references is repeated in the body of the report.
General	Suggested Citation, Preface, and Acknowledgements are absent from this draft.	We have added these sections to the final draft.
Appendices	Alphabetical order for TEP members please	The requested change has been made.
Appendices	There were multiple calls prior to the EPC getting involved; should specify which call and the date the TEP member participated in.	We have added the requested information.
Appendices	p. 2 appendices delete “clinical expert” replace with “surgery”	The requested change has been made.
Appendices	Need to replace Vienna with Austria.	The requested change has been made.