Noninvasive Diagnostic Tests for Breast Abnormalities: Update of a 2006 Review

Executive Summary

Background

Breast cancer is one of the most common malignancies of women, with approximately 200,000 new cases diagnosed every year in the United States.\(^1\) Some breast cancers are identified by physical examination (either self-examination or an examination performed by a physician). Population-wide screening programs that use x-ray mammography to examine asymptomatic women for early signs of breast cancer are also in common use.\(^2\)-\(^4\) If a suspicious area is seen on x-ray mammography, women are usually recalled for further examination. The results of these examinations are used to make decisions about further management: return to normal screening/return for short-interval followup/refer for biopsy. In current standard practice the examinations conducted after recall usually consist of diagnostic mammography and possibly ultrasound. More and more often women are being sent for additional imaging during recall workup. Extensive diagnostic ultrasound examinations and MRI are currently the most commonly chosen additional imaging added to the workup, but other imaging technologies are offered by some practitioners.

It is important to triage recalled women into the correct management pathway.

Women with readily treatable early-stage cancers who mistakenly get triaged into “return to normal screening” may experience a significant delay in diagnosis and treatment of the cancer. However, the majority of women who are recalled for
Further assessment after a screening mammography do not have cancer, and significant numbers of healthy women are referred for biopsy or short-interval followup after recall and diagnostic mammography.5,6

A number of noninvasive imaging technologies have been developed and proposed to be useful as part of the workup after recall. This evidence review focuses on additional noninvasive imaging studies that can be conducted (in addition to standard workup) after discovery of a possible abnormality on screening mammography or physical examination. These studies are intended to guide patient management decisions. In other words, these imaging studies are not intended to provide a final diagnosis as to the nature of the breast lesion; rather, they are intended to provide additional information about the nature of the lesions such that women can be more appropriately triaged into the correct management pathway. It is important to evaluate the evidence to see if women do or do not benefit from the addition of these imaging modalities to the standard workup after recall on breast cancer screening.

Because there are no available studies that directly evaluate whether women benefit from additional imaging in this context, we addressed this important question indirectly. First we evaluated the accuracy of the imaging tests in distinguishing between “benign” and “malignant” breast lesions. Inaccurate tests will lead to suboptimal management decisions and less than desirable patient outcomes. The accuracy of the noninvasive imaging tests was primarily measured in terms of sensitivity and specificity. Sensitivity is a measure of how accurately the test can identify women with cancer; specificity is a measure of how accurately the test can identify women who do not have cancer. A test with high sensitivity will rarely missclassify women with cancer as not having cancer, and a test with high specificity will rarely misclassify women without cancer as having cancer.

The accuracy of a test can also be expressed in a more clinically useful measure, namely, likelihood ratios. When making medical decisions, a clinician can use likelihood ratios and test results to estimate the probability of an individual woman having breast cancer. Clinicians use individual patient characteristics (such as age and family history) and features seen on the diagnostic mammogram (such as microcalcifications or distortions) to estimate a woman’s risk of malignancy. This estimate is known as a “pre-test” or “prior” probability. The clinician can then use the likelihood ratios (that express the accuracy of the test) to decide if an additional imaging test will be helpful in guiding management decisions. For example, if a clinician estimates a woman’s risk of malignancy as greater than 50 percent, most likely the use of any additional imaging test, even a very accurate imaging test, will not change the clinician’s management recommendation of a biopsy, and therefore additional imaging will not be beneficial to the woman. However, if a clinician estimates a woman’s risk of malignancy as being uncertain or close to a clinical threshold (2%), the likelihood ratios can be used to estimate whether the results of an additional test are likely to change management decisions and possibly affect patient outcomes.

After establishing the accuracy of the various imaging tests, we used the summary likelihood ratios to prepare simple models of various clinical scenarios. In doing so, we attempted to indirectly address the implicit question of whether women benefit from the addition of noninvasive imaging tests to standard workup after recall for evaluation of a possible breast abnormality detected by screening mammography or physical examination.

This report is an update of a Comparative Effectiveness Review (CER) of the same title originally published in 2006.7 In addition to an update of the literature, the Key Questions have been revised and additional noninvasive imaging tests have been added.

Methods

Topic Development and Scope

The topic was selected for update by the Effective Health Care program. The Key Questions were posted for public comment. A Technical Expert Panel was assembled to provide expert input, and a protocol for updating the review was developed by the EPC authors and approved by the Agency for Healthcare Research and Quality.

Patient Population

The patient population of interest is the general population of women participating in routine breast cancer screening programs (including mammography, clinical examination, and self-examination) who have been recalled after discovery of a possible abnormality and who have already undergone standard workup (which usually includes diagnostic mammography and/or ultrasound). In other words, the patient population of interest consists of women who have or might receive a Breast Imaging-Reporting and Data System (BI-RADS®) rating of 0, or 3 to 5, after standard workup. Some of the women evaluated may have had an ultrasound examination before being examined using the technology under study, including the women being evaluated by diagnostic ultrasound. Although not explicitly stated in the studies, in most cases this prior
ultrasound seemed to be used primarily to identify women with simple benign cysts, who were then not included in the study. Populations that were not evaluated in this review include: women thought to be at very high risk of breast cancer due to family history or breast cancer (BRCA) gene mutations; women with a personal history of breast cancer; women presenting with overt symptoms (such as pain or nipple discharge); and men.

**Interventions**
The noninvasive diagnostic tests evaluated were ultrasound (conventional B mode grayscale, harmonic, tomography, color Doppler, and power Doppler); magnetic resonance imaging (MRI, with gadolinium-based contrast agents) with or without computer-aided diagnosis (CADx); positron emission tomography (PET, with 18-fluorodeoxyglucose [FDG]), with or without concurrent computed tomography (CT) scans (including positron emission mammography [PEM]); scintimammography (with technetium-99m sestamibi [MIBI]), including Breast Specific Gamma Imaging (BSGI).

**Comparators**
The accuracy of the noninvasive diagnostic tests were evaluated by a direct comparison with histopathology (surgical or biopsy specimens) or with clinical followup, or a combination of these methods. In addition, the relative accuracy of the different tests under evaluation were directly and indirectly compared as the evidence permitted.

**Outcomes**
Outcomes of interest are diagnostic test characteristics; namely, sensitivity, specificity, and likelihood ratios. Because predictive values vary as the prevalence of disease changes, we did not calculate predictive values. Adverse events related to the procedures, such as radiation exposure, discomfort, and reactions to contrast agents, were also be discussed as the evidence permitted. Our literature searches did not identify any relevant studies that directly reported the impact of the diagnostic tests on patient-oriented outcomes. Therefore, we used the estimates of accuracy and various clinical scenarios to address the implicit, very important question of whether women benefit from the use of these noninvasive imaging tests.

**Timing**
Any duration of followup, from same-day interventions to many years of clinical followup, were evaluated.

**Setting**
Any care setting was evaluated, including general hospitals, physician’s offices, and specialized breast imaging centers.

**Study Selection**
We searched the medical literature, including PubMed and Embase, from December 1994 through September 2010. We included diagnostic cohort studies that enrolled the patient population of interest and used current generation scanners and protocols of the noninvasive imaging technologies of interest. We excluded case-control studies, meeting presentations, and very small (<10 patients) studies. Data were abstracted from the included studies.

**Strength of Evidence**
We graded the strength of evidence supporting each major conclusion as high, moderate, low, or insufficient. The grade was developed by considering four important domains: the risk of bias in the evidence base (internal validity, or the quality of the studies), the consistency of the findings, the precision of the results, and the directness of the evidence.

**Data Analysis**
We used a bivariate mixed-effects binomial regression model for meta-analysis of data.\(^8,9\) We used summary likelihood ratios and Bayes’ theorem to calculate the post-test probability of having a benign or malignant lesion. In cases where a bivariate binomial model could not be fit, we meta-analyzed the data using two random-effects models, one for sensitivity and one for specificity.\(^10\) We explored heterogeneity in the data with meta-regressions using standard methodology.\(^9\)

**Peer Review and Public Commentary**
The draft received comments from peer reviewers, and from members of the public through an open public comment period.

**Results**

**Magnetic Resonance Imaging**
We identified 41 studies of MRI that included a total of 3,882 patients with 4,202 suspicious breast lesions.\(^11-51\) We combined the data reported by all 41 studies into a bivariate binomial mixed-effects model. The summary sensitivity was 91.7 percent (95% CI: 88.5 to 94.1%)
and the summary specificity was 77.5 percent (95% CI: 71.0 to 82.9%). The estimate of accuracy was judged to be supported by a moderate to low strength of evidence (low for the estimate of specificity due to the wide confidence interval). The dataset was very heterogeneous ($I^2 = 98.4$%). We explored the heterogeneity with meta-regression and found that the prevalence of disease in the study population and whether or not the image readers were blinded was statistically significantly correlated with the results. Subgroup analyses found that MRI was less sensitive for evaluation of microcalcifications (84.0% vs. 91.7% summary sensitivity).

The probability that a woman actually has cancer (invasive or in situ) even after a finding of “benign” on MRI depends on her probability of having cancer before undergoing the test. Bayes’ theorem and the summary likelihood ratios indicate that if a woman with an estimated 5 to 10 percent chance of having cancer undergoes MRI and has a finding of “benign,” she will then have an estimated 1 percent chance of having cancer; a woman with an estimated 20 percent chance of having cancer who has a finding of “benign” on MRI will then have an estimated 3 percent chance of having cancer; and a woman with an estimated 50 percent chance of having cancer who has a finding of “benign” on MRI will then have an estimated 10 percent chance of having cancer.

**Positron Emission Tomography**

We identified seven studies of PET$^{34,35,41,52-55}$ and one study of PET/CT$^{16}$ that met our inclusion criteria. The studies included a total of 308 women with 403 suspicious breast lesions. We combined the data reported by the seven studies of PET into a bivariate binomial mixed-effects model. The summary sensitivity was 83.0 percent (95% CI: 73.0 to 89.0%) and the summary specificity was 74.0 percent (95% CI: 58.0 to 86.0%). The estimate of accuracy was judged to be supported by a Low strength of evidence. The dataset contained moderate heterogeneity ($I^2 = 64.0$%). We explored the heterogeneity with meta-regression and did not identify any possible causes. Subgroup analyses found that PET was more sensitive for evaluation of palpable lesions.

The probability that a woman actually does have cancer (invasive or in situ) even after a finding of “benign” on PET depends on her probability of having cancer before undergoing the test. Bayes’ theorem and the summary likelihood ratios indicate that if a woman with an estimated 5 percent chance of having cancer undergoes PET and has a finding of “benign” she will then have an estimated 1 percent chance of having cancer; a woman with an estimated 20 percent chance of having cancer who has a finding of “benign” on PET will then have an estimated 6 percent chance of having cancer; and a woman with an estimated 50 percent chance of having cancer who has a finding of “benign” on PET will then have an estimated 19 percent chance of having cancer.

**Scintimammography**

We identified 10 studies of scintimammography$^{14,56-64}$ and one study of BSGI$^{19}$ that met our inclusion criteria. The studies included a total of 1,064 suspicious lesions. We combined the data reported by all 11 studies into a bivariate binomial mixed-effects model. The summary sensitivity was 84.7 percent (95% CI: 78.0 to 89.7%) and the summary specificity was 77.0 percent (95% CI: 64.7 to 85.9%). The estimate of accuracy was judged to be supported by a low strength of evidence. The dataset was very heterogeneous ($I^2 = 93.0$%). We explored the heterogeneity with meta-regression and did not identify any possible causes.

The probability that a woman actually does have cancer (invasive or in situ) even after a finding of “benign” on scintimammography depends on her probability of having cancer before undergoing the test. Bayes’ theorem and the summary likelihood ratios indicate that if a woman with an estimated 5 percent chance of having cancer undergoes scintimammography and has a finding of “benign,” she will then have an estimated 1 percent chance of having cancer; a woman with an estimated 20 percent chance of having cancer who has a finding of “benign” on scintimammography will then have an estimated 5 percent chance of having cancer; and a woman with an estimated 50 percent chance of having cancer who has a finding of “benign” on scintimammography will then have an estimated 17 percent chance of having cancer.

**Ultrasound**

We identified a total of 31 diagnostic cohort studies of ultrasound. Of these, there were 21 studies of B-mode grayscale ultrasound,$^{18,26,65-83}$ six studies of color Doppler ultrasound,$^{78,80,84-87}$ and nine studies of power Doppler ultrasound.$^{65,72,75,77,86,88-91}$ We combined the data reported by these studies into bivariate binomial mixed-effects models. For B-mode grayscale, summary sensitivity was 92.4 percent (95% CI: 84.6 to 96.4%) and the summary specificity was 75.8 percent (95% CI: 60.8 to 86.3%); for color Doppler, summary sensitivity was 88.5 percent (95% CI: 74.4 to 95.4%) and summary specificity was 92.4 percent (95% CI: 84.6 to 96.4%); for power Doppler, summary sensitivity was 70.8 percent (95% CI: 47 to
86.6%) and summary specificity was 72.6 percent (95% CI: 59.9 to 82.5%). These estimates of accuracy were all judged to be supported by a low strength of evidence. The datasets were heterogeneous. We explored the heterogeneity of the largest dataset (21 studies of B-mode) with meta-regression and found that whether the studies blinded the image readers and accounted for inter-reader differences were statistically significantly associated with the results.

The probability that a woman actually does have cancer (invasive or in situ) even after a finding of “benign” on ultrasound depends on her probability of having cancer before undergoing the test. Bayes’ theorem and the summary likelihood ratios indicate that if a woman with an estimated 5 to 10 percent chance of having cancer undergoes B-mode grayscale ultrasound and has a finding of “benign,” she will then have an estimated 1 percent chance of having cancer; a woman with an estimated 20 percent chance of having cancer who has a finding of “benign” on B-mode grayscale ultrasound will then have an estimated 2 percent chance of having cancer; and a woman with an estimated 50 percent chance of having cancer who has a finding of “benign” on B-mode grayscale ultrasound will then have an estimated 9 percent chance of having cancer.

**Discussion**

According to the American College of Radiology, the threshold of suspicion of malignancy at which management of women changes is 2 percent. After recall and workup, women with a suspicion of malignancy greater than 2 percent are generally recommended to undergo tissue sampling of some kind (biopsy), and women with a lower suspicion of malignancy are triaged into imaging management pathways (short-interval followup or return to regular screening). We used the 2 percent threshold to explore the clinical usefulness of the various noninvasive imaging technologies as add-ons to the current standard of care; namely, if a woman was recalled for evaluation after a screening mammography, and received standard-of-care workup versus standard-of-care workup plus the noninvasive imaging technology, would use of the noninvasive imaging technology be likely to alter the recommendations for care after the workup?

For all of the technologies evaluated in this assessment, only women with a low suspicion of malignancy after standard-of-care workup might be expected to experience a change in management decisions as a result of additional noninvasive imaging. A woman with a ≤12 percent suspicion of malignancy who has benign findings on MRI could have her suspicion of malignancy drop below the 2 percent threshold, and therefore she might be assigned to short-interval imaging followup management rather than tissue sampling management; a woman with a 1 percent suspicion of malignancy who has benign findings on MRI could have her suspicion of malignancy drop to near 0 percent, and therefore she might be assigned to return to normal screening rather than short-interval followup imaging; a woman with a 1 percent suspicion of malignancy who has malignant findings on MRI could have her suspicion of malignancy increase to 4 percent, and therefore she might be assigned to tissue sampling management rather than short-interval followup. The equivalent thresholds of pretest suspicion of malignancy at which additional imaging may change management are: for B-mode grayscale ultrasound, 1 to 10 percent; for scintimammography, 1 to 5 percent; and for PET, 1 to 5 percent.

Therefore, if the 2 percent threshold is chosen, the use of noninvasive imaging in addition to standard workup may be clinically useful for diagnostic purposes only for women with a low suspicion of malignancy. When choosing which noninvasive imaging technology to use for this purpose, diagnostic B-mode grayscale ultrasound and MRI appear to be more accurate than PET, scintimammography, or the other types of ultrasound (e.g., Doppler) that were evaluated in this comparative effectiveness review.

Women thought to be at moderate to high risk of malignancy after standard workup will not have their estimate of risk of malignancy change sufficiently after further noninvasive imaging to affect management decisions. For many patients the suspicion of malignancy will not be able to be estimated with sufficient precision for clinicians to feel comfortable recommending return to normal screening (rather than a biopsy or short-interval followup) solely on the basis of additional noninvasive imaging. Estimates of risk of malignancy are based on features of the mammographic images, patient characteristics, patient history, and patient family history. Several of our expert reviewers did not think such precise estimation of risk is feasible using currently available methods. Potential harms of noninvasive imaging, such as radiation exposure, also need to be considered when deciding whether to perform these tests.
Changes Since 2006

This CER is an update of a CER finalized in 2006. The updated results are, in general, very similar to the findings of the 2006 report. For MRI, in 2006 we found that the sensitivity was 92.5 percent and the specificity was 75.5 percent; the updated evidence base supported estimates of 91.7 percent sensitivity and 77.5 percent specificity. In both reports, MRI was found to be less sensitive (approximately 85%) for evaluation of microcalcifications than for evaluation of lesions in general. For PET, in 2006 we found that the sensitivity was 82.2 percent and the specificity was 78.3 percent; the updated evidence base supported estimates of 83.0 percent sensitivity and 74.0 percent specificity. In the updated report we attempted to evaluate the accuracy of PET/CT, but only one study that met the inclusion criteria was identified.

For scintimammography, the updated evidence base identified a sensitivity of 84.7 percent, much higher than the sensitivity estimate from 2006 of 68.7 percent. Specificity was estimated at 84.8 percent in 2006, and at 77.0 percent in the update; however, the confidence intervals around the updated estimate of specificity are wide. It is possible that improvements in the technology in the last few years improved the sensitivity of the technique.

For ultrasound, in 2006 we evaluated a relatively small set of studies of B-mode grayscale ultrasound, and estimated a sensitivity of 86.1 percent and a specificity of 66.4 percent. The update included a significantly expanded evidence base on B-mode grayscale ultrasound, and identified a sensitivity of 92.4 percent and specificity of 75.8 percent. In the update we included numerous other types of ultrasound, including power and color Doppler ultrasound, that were not studied in the 2006 report.

Remaining Issues

The conclusions of quantitative accuracy were for the most part rated as being supported by low strength of evidence, due primarily to the imprecision of the estimates (wide confidence intervals around the estimates of accuracy); the publication of additional diagnostic accuracy studies are likely to increase the precision of the estimates of accuracy, which may upgrade the strength of evidence rating. There was also considerable heterogeneity (inconsistency) in the majority of the evidence bases, which contributed to the low strength of evidence rating. Most likely the heterogeneity was due to slight differences in imaging methodology or patient populations across studies; future research intended to tease out factors affecting the accuracy of imaging may be helpful to the clinician when deciding whether a test may be a useful addition to standard workup for management of a particular patient.

However, the publication of additional diagnostic accuracy studies is unlikely to affect the implications of the conclusions. The conclusions of diagnostic accuracy lead indirectly to a conclusion that only women with a low (1 to 12%) suspicion of malignancy will experience a “change in management” (which may or may not be beneficial) from the use of these noninvasive diagnostic tests. Improving the precision of the estimates of accuracy or upgrading the strength of evidence rating in response to the publication of more diagnostic accuracy studies will not affect the indirect conclusion. Studies that address the issue of how to establish more accurate estimates of malignancy from diagnostic mammography for an individual patient may be more clinically relevant than additional diagnostic accuracy studies.

A limitation of the current evidence base that should be addressed in future research is the patient population being evaluated. Many of the currently available studies were conducted only on women who had been scheduled for biopsy after standard workup, and therefore the patient population studied is not truly representative of the entire patient population of interest. Additional studies that enroll women referred for short-interval followup after standard workup are needed to confirm that the findings of this assessment do apply to the patient population of interest.

In addition, the majority of studies did not report data separately for different categories of breast lesions or patient characteristics. Future research should focus on the accuracy of noninvasive imaging technologies for discrete categories of lesions, such as nonpalpable lesions classified as BI-RADS 3, or for discrete categories of women, such as women older than age 75. Information from more granular groupings of women will allow estimates of test accuracy to be more immediately clinically useful.

Future research efforts should also focus on studies that report the impact of the use of noninvasive imaging on patient-oriented outcomes such as quality of life, and on evaluation of newer noninvasive imaging technologies.

Conclusions

Our key findings are summarized in Table A. In conclusion, the use of noninvasive imaging in addition to standard workup after recall for evaluation of a breast lesion detected on screening mammography or physical
examination may be clinically useful for diagnostic purposes only for women with a low (1 to 12%) suspicion of malignancy. When choosing which noninvasive imaging technology to use for this purpose, diagnostic B-mode grayscale ultrasound and MRI appear to more accurate than PET, scintimammography, or Doppler ultrasound. However, whether these findings are clinically relevant hinges on whether clinicians can identify those women who, after standard workup after recall, have a risk of malignancy in this range. Several expert reviewers of this report expressed doubt about the feasibility of such precise estimation.

### Table A. Summary of key findings

<table>
<thead>
<tr>
<th>Technology</th>
<th>Summary Sensitivity</th>
<th>Summary Specificity</th>
<th>Pretest Probability of Malignancy Threshold&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Strength of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-mode grayscale 2D ultrasound</td>
<td>92.4% (84.6 to 96.4%)</td>
<td>75.8% (60.8 to 86.3%)</td>
<td>1 to 10%</td>
<td>Low</td>
</tr>
<tr>
<td>MRI</td>
<td>91.7% (88.5 to 94.1%)</td>
<td>77.5% (71.0 to 82.9%)</td>
<td>1 to 12%</td>
<td>Moderate (sensitivity) to Low (specificity)</td>
</tr>
<tr>
<td>Scintimammography</td>
<td>84.7% (78.0 to 89.7%)</td>
<td>77.0% (64.7 to 85.9%)</td>
<td>1 to 5%</td>
<td>Low</td>
</tr>
<tr>
<td>PET</td>
<td>83.0% (73.0 to 89.0%)</td>
<td>74.0% (58.0 to 86%)</td>
<td>1 to 5%</td>
<td>Low</td>
</tr>
</tbody>
</table>

<sup>a</sup>The threshold at which use of the noninvasive imaging test may change the post-test probability of malignancy sufficiently to trigger a change in patient management.

### References


Full Report

This executive summary is part of the following document:

For More Copies

For more copies of Noninvasive Diagnostic Tests for Breast Abnormalities: Update of a 2006 Review: Comparative Effectiveness Review Executive Summary No. 47 (AHRQ Pub. No. 12-EHC014-1), please call the AHRQ Clearinghouse at 1-800-358-9295 or email ahrqpubs@ahrq.gov.