Published in final edited form as:

Acad Radiol. 2017 November ; 24(11): 1436–1446. doi:10.1016/j.acra.2017.05.007.

The reproducibility of changes in diagnostic figures of merit across laboratory and clinical imaging reader studies

Frank W. Samuelson and

U.S. Food and Drug Administration, 10903 New Hampshire Ave., Building 62, Room 3102, Silver Spring, Maryland 20993-0002

Craig K. Abbey

Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106

Abstract

Rationale and Objectives—In this paper we examine which comparisons of reading performance between diagnostic imaging systems made in controlled retrospective laboratory studies may be representative of what we observe in later clinical studies. The change in a meaningful diagnostic figure of merit between two diagnostic modalities should be qualitatively or quantitatively comparable across all kinds of studies.

Materials and Methods—In this meta-study we examine the reproducibility of relative sensitivity, false positive fraction, area under the ROC curve, and expected utility across laboratory and observational clinical studies for several different breast imaging modalities, including screen film mammography, digital mammography, breast tomosynthesis, and ultrasound.

Results—Across studies of all types the changes in the false positive fractions yielded very small probabilities of having a common mean value. The probabilities of relative sensitivity being the same across ultrasound and tomosynthesis studies were low. No evidence was found for different mean values of relative area under the ROC curve or relative expected utility within any of the study sets.

Conclusion—The comparison demonstrates that the ratios of areas under the ROC curve and expected utilities are reproducible across laboratory and clinical studies, whereas sensitivity and false positive fraction are not.

Keywords

sensitivity; specificity; AUC; reproducibility

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Studies of reader performance play an important role in evaluations of the effectiveness of new diagnostic imaging devices and methodologies [1, 2, 3, 4, 5]. These studies are particularly common in applications where a diagnostic task can be simplified to a binary choice, as in breast cancer screening where the goal is to identify abnormalities for subsequent analysis. From this perspective, the screening exam sorts the cases into “negative” and “positive” categories. Binary tasks are commonly characterized by a receiver operating characteristic (ROC) curve, which plots true positive fraction (TPF; the fraction of disease cases correctly labeled positive) as a function of the false positive fraction (FPF; the fraction of non-disease cases incorrectly labeled as positive). The ROC curve serves as the underlying framework from which various figures of merit (FOMs) of imaging performance can be defined.

Before utilizing a new diagnostic imaging modality in clinical practice, a laboratory reader study of the modality may be performed to determine its diagnostic effectiveness for regulatory purposes or to demonstrate its capabilities for the medical community. Often this study will compare the performance of the new modality in one study arm to the standard of care in a control arm [1, 4, 6, 7, 8]. These pre-clinical reader studies differ from clinical trials in a number of ways that make the studies far less time-consuming and costly. They generally require a much smaller sample of patient exams, which limits the potential for side-effects of imaging including exposure to ionizing radiation or intravenous contrast agents. In imaging tasks that have low disease prevalence, e.g. asymptomatic breast-cancer screening, the patient sample is usually enriched with cases of disease through various possible sampling strategies [9]. Other differences may include limited access to additional clinical data (prior exams, patient symptoms, or history), a different reporting format that may include non-standard measures such as a probability of malignancy rating, and use of retrospective data which results in a lack of direct consequences for patient management.

After initial acceptance of an imaging modality, clinical observation studies are often reported as a way to provide an assessment of the new modality in practice. These studies typically compare the new modality to the pre-existing standard in a clinical setting with a clinical population and with patient management decisions based on outcomes from imaging. Observational studies are instrumental for determining the success of a new imaging modality in the health care environment, for making reimbursement decisions, and potentially for defining a new standard of care.

The literature on pre-clinical laboratory studies includes an extended debate over what should be the appropriate figure of merit (FOM) for quantifying diagnostic performance of imaging systems [5, 2, 10, 11, 12, 13, 14, 15, 16]. Proposed figures of merit include TPF and FPF [17, 18, 19], area under the ROC curve (AUC) [20, 21, 22], and utility measures [23, 24, 25, 26, 27] among others [28, 29, 30, 31]. TPF and FPF are considered criterion-dependent measures because they result from a perceived signal exceeding a critical value in classical signal detection models. They are understood in these models as an operating point on an ROC curve. In contrast, AUC is considered criterion-free because it does not depend on any one operating point. We also note that there are other approaches to defining imaging

performance based on assessing localization accuracy [32, 33, 34] with a similar debate over FOMs [35, 36, 37] that this work will not address.

Much of the debate has focused on how a FOM may be interpreted as a measure of performance, or what advantages a FOM may have in experimental design such as statistical power. A consideration that has received far less attention is the issue of reproducibility. A most fundamental feature of any FOM is its ability to measure differences between experimental conditions reproducibly. If a FOM cannot be reproduced quantitatively, or at least qualitatively, across scientific studies, then it is not useful. Reproducibility is the focus of this work.

Much recent literature has been devoted to the importance of the reproducibility of scientific results [38], particularly in medicine. In this work we examine the reproducibility of different FOMs across initial laboratory studies and how well those FOMs translate to the subsequent clinical observational studies. Specifically we are interested in which FOMs demonstrate qualitative or quantitative reproducibility of measured changes between the experimental and control arms of the studies. In this paper we use a broad definition of the term reproducibility to mean how well a FOM is reproduced across both controlled and observational studies with different designs and settings.

We compare several published laboratory studies on screen film mammography (SFM), full field digital mammography (FFDM), breast ultrasound imaging (US), and digital breast tomosynthesis (DM+DBT) with the published observational studies that followed. These studies examine imaging technologies used to screen patients for breast cancer. We focus on imaging modalities related to breast cancer screening because these have received considerable interest over time, and hence there are more published data available.

There are a number of technical challenges when computing the FOMs from the data reported in observational studies. A primary difficulty is that many large observation clinical studies do not follow all patients who are called negative by the diagnostic test. Some of these patients will truly be diseased, and the number of false negative patients and the true number of diseased patients in those studies are unknown. Often these observational studies only report rates of detection and recall. From these studies we can not directly calculate some FOMs, like true or false positive fractions, or their differences. However we can calculate or approximate the relative values of these statistics between study arms when the arms use the same patient population [39, 40]. Therefore we quantify effects in terms of ratios of FOMs of a new modality to a standard modality.

2. Materials and Methods

The true positive fraction, TPF, is the fraction of all diseased patients that are classified as positive by the diagnostic. Likewise the false positive fraction, FPF, is the fraction of all non-diseased patients who are classified as positive by the diagnostic. Specificity or the true negative fraction is simply the complement of the FPF, $TNF=1-FPF$, so an increase in FPF requires a decrease in specificity. Therefore any inferences we make regarding the

reproducibility of changes in FPF also apply to specificity. TPF and FPF are easily computed, so they are frequently reported in diagnostic imaging studies.

Most diagnostics do not just have positive or negative outputs. Diagnostics measure a concentration like a bio-marker, or an X-ray density like a scanner, or a reader's confidence of disease. Typically if that measured value for a patient is greater than some threshold, we may consider that a positive diagnosis. Values below the threshold are interpreted as negative. However that threshold can be changed, and as we change that threshold TPF and FPF also change. How TPF and FPF vary with respect to each other as the decision threshold changes is the ROC curve [20], an example of which is shown in Figure 1.

The area under this ROC curve (AUC) is frequently used as a summary measure of diagnostic accuracy. The AUC can be interpreted many ways. It is the average sensitivity (TPF) over all specificities or the average specificity (TNF) over all sensitivities. It is also the probability that a diagnostic will correctly identify which of two randomly selected patients is diseased [7]. AUC is frequently used as a measure of overall diagnostic accuracy.

“Expected utility” [41] is a measure of the trade-off between the true positive fraction and the false positive fraction at a clinical decision threshold. Specifically, it is defined as

$$EU = TPF_c - \beta \times FPF_c \quad (1)$$

where TPF_c and FPF_c are the true and false positive fractions on a reader's ROC curve where the slope of the curve is β . Based on the clinical practice of mammography Abbey et al. [26] determined that the average value of β is approximately 1.03 for breast cancer screening in the United States. Figure 1 shows EU graphically as an intercept of a line with slope β that is tangent to the diagnostic test's ROC curve at the optimal or clinical TPF, FPF operating point.

2.1. Relative statistics

We know that readers may change their decision thresholds or confidence scores based on the perceived prevalence of the targets and the costs and benefits of calling them [42, 43]. Therefore FOMs such as TPF and FPF, which depend upon reader decision thresholds, may be different between studies that have different disease prevalence or different costs or benefits [44], such as between retrospective laboratory studies and observational clinical studies.

FOMs such as AUC are not dependent on decision thresholds, and therefore independent of perceived prevalence or costs [3]. However, their absolute values may still differ between laboratory studies and clinical studies because the laboratory studies may contain a more difficult (or easier) spectrum of cases, depending on how they were selected.

Fortunately most medical imaging studies provide two arms, an arm with the new experimental modality and a control arm of standard practice, both using the same patients and readers, or both using the same population of patients and readers. We expect that the

change in an FOM between the study arms in laboratory studies will be representative of the change in later clinical studies, either qualitatively or quantitatively. For example, if FOM X measured in a controlled laboratory study tells us that diagnostic imaging modality A is 30% better than modality B, then FOM X should qualitatively demonstrate an increase between the same modalities in a large observational clinical trial. Preferably the increase also should be quantitatively similar.

To evaluate the change in FOMs between modalities in our study sample we used relative statistics. We calculated the point estimates and the standard errors of the average relative true positive fraction (rTPF), relative false positive rate (rFPF), relative AUC (rAUC), and relative expected utility (rEU) for each study. Relative statistics are the ratios of the metric between the two arms of the study. The relative TPF is the ratio of the average TPFs in each arm of the study, $rTPF = TPF_A / TPF_B$. A rTPF that is significantly greater than one indicates that modality A has higher sensitivity than modality B.

Frequently studies use differences of FOMs to infer changes between study arms, i.e. $TPF_A - TPF_B$, but we used relative statistics because they are independent or insensitive to the unknown false negative fractions or disease prevalence as discussed below. Inferences with relative statistics are the same as inferences with differences when the measured values are positive. Additionally relative changes in FOMs may be meaningful across studies where absolute changes are not. For example, in a laboratory study by Rafferty et al. [45] the false positive rate reduced by 0.19 or 38%. In clinical studies of breast cancer screening false positive fractions are typically around 0.10, so an absolute reduction of 0.19 is not possible, but a 38% decrease of FPF to 0.62 may be possible.

Some of the studies that we reviewed had follow-up or biopsy data on all the patients in the study, and therefore they reported true positive fractions and false positive fractions. These studies are indicated using a “C” in the tables. We used the fractions that were reported at the “recall” threshold for mammography where possible.

Most of the clinical studies reported average rates or fractions (TPF, FPF, recall or detection) as the number of all patients called positive by all readers divided by the relevant number of patients. In other words the reader's evaluations were pooled. This is in contrast to averaging the fractions calculated for each reader, which is often done in laboratory reader studies. For studies where such data were available, we calculated the average fractions in both manners and found that it made little difference in the point estimates.

Many of the laboratory studies and some of the clinical studies reported AUC values for both arms of the study. These studies are indicated using a “D” in the tables. The ratio of these AUC values was our relative AUC value, rAUC. Some of these AUC estimates were semi-parametric, some were non-parametric. For studies that did not report AUC values, we used ordinal regression [46, 47] to create semi-parametric power-law estimates [48, 49] of the ROC curves given what TPF and FPF values were provided to us or in the study paper. In many clinical studies we had only a single TPF/FPF measurement on which to base our semi-parametric ROC curve and AUC. In general we do not recommend calculating AUC based on a single decision threshold unless no other data are available.

To calculate the relative expected utility, rEU, in all the clinical studies in our sample we utilized equation 1 with $\beta = 1.03$ and the same reported true and false positive fractions that we used to calculate rTPF and rFPF. The ratio of two EU values can also be calculated directly from the recall and detection fractions cited in observational clinical studies [40]. This calculation assumes that the readers in these studies were operating with the same or similar clinical utilities as those studied by Abbey et al. [26]. We assumed equal clinical utility because all the studies in our sample were breast cancer screening studies. Some of the clinical studies were performed in Europe or on populations of patients with only dense breasts, whereas the derivation of the average β trade-off value was calculated from the general population in the United States [26]. We did not attempt to adjust β for these situations.

For all laboratory reader studies we did not assume that readers were operating near a clinical decision threshold, because high disease prevalence strongly affects true and false positive fractions of readers [44]. To calculate EU in each laboratory study we estimated a semi-parametric ROC curve, as with AUC, and calculated the expected utility at the point on the ROC curve where the slope was $\beta = 1.03$. For an example of determining EU from an ROC curve, see Abbey et al. [27]. Therefore the calculation of EU in clinical studies was analytically different than in the laboratory studies. However we expected a priori that all these calculations were measuring the same quantity.

2.2. Relative statistics in studies with partial disease confirmation

As noted in the introduction, many observational clinical studies only collect binary (positive or negative) outcomes from radiologist evaluations of patient images, and many of these studies do not confirm the disease status of patients who are called negative by the diagnostic imaging devices under study. Therefore estimating the true prevalence of disease π in the population of the study is not possible. Likewise, without assuming a disease prevalence we can not estimate a true or false positive fraction that we can compare with other studies, such as controlled laboratory studies. Observational clinical studies frequently report recall rates (the fraction of all patients called positive by the diagnostic, R) and detection rates (the fraction of all patients called positive by the diagnostic and verified to have the disease, D), which can be calculated without knowing the number of false negatives or number of verified diseased patients. If a clinical study compares the performance of two imaging diagnostics on the same patients or the same population, then the prevalence of diseased patients is the same for both study arms, or the expected prevalence is the same. In that scenario, ratios of some diagnostics statistics can be calculated without dependence on the prevalence or the number of diseased patients who were called negative by the diagnostic tests [39].

If we knew the total number of patients in a study who were actually diseased (M), then we would estimate the sensitivity or true positive fraction of diagnostic A as $TPF_A = TP_A/M$, where (TP_A) is the number of true positive patients. In many observational clinical studies M is not known, and therefore we can not calculate an absolute TPF or a difference in TPFs. If a study measured the number of true positives from two diagnostics on the same diseased patients, then the ratio of those true positives is the ratio of the true positive fractions.

$$rTPF = \frac{TPF_A}{TPF_B} = \frac{TP_A/M}{TP_B/M} = \frac{TP_A}{TP_B}$$

Therefore we can estimate the relative true positive fraction as the ratio of the number of true positives, without knowing the actual number of diseased patients. This same principle applies to the ratio of two false positive fractions [39] or two expected utilities [40]. The relative false positive fraction is calculated as the ratio of the number of false positives for each diagnostic, $rFPF = FP_A/FP_B$, where FP_A is the number of all recalls minus the number of cancer detections in study arm A. When sample sizes in the study arms are unequal, these ratios scale by the ratio of the total numbers of patients in each study arm.

In addition to FOMs like rTPF, which is independent of the number of diseased patients or the disease prevalence, there are relative FOMs such as rAUC that can be estimated from the positive rates measured in observational clinical studies with only weak dependence on the unknown disease prevalence. When the differences in recall and detection rates between arms of a study are not large, then rAUC has a weak, second-order dependence on the posited prevalence [40]. This weak dependence of rAUC on prevalence holds for almost all of the clinical studies examined in this paper.

For those clinical studies that did not report number or fraction of verified diseased patients, we assumed values of disease prevalence π that were larger than the observed detection rates in the studies, but still consistent with other published studies [50, 51]. Those assumed values are given in parenthesis in Tables 2 and 3. From those posited fractions of diseased patients we calculated approximate TPF and FPF values and fit a semi-parametric ROC model to these TPF, FPF values for each arm of the study. The ratio of the areas under those ROC curves is our estimate of rAUC. The systematic errors in rAUC due to our uncertainty in prevalence are less than the statistical errors in almost all the observational clinical studies. Therefore even if we have studies where the prevalence of disease or the conditions of negative patients are unknown, we can compare relative diagnostic FOMs among studies, and those relative FOMs should be similar.

2.3. Estimating variability in imaging studies

Most of the studies that we reviewed reported estimates of diagnostic FOMs and their standard errors or confidence intervals. For most studies we used the error reported for each modality and the error on their difference to calculate an error on the ratio of FOMs using the δ -method [52, 53]. For several studies the errors on the modalities or their differences had to be inferred from quoted confidence intervals or tabulated data.

Some studies, such as Kerlikowske et al. [54] correctly estimated statistical errors using sites and radiologists as random effects as well as patients as random effects. The reported errors of Kerlikowske et al. are much larger than binomial errors given patient sample sizes. Many other observational clinical studies did not correctly estimate errors in this way. These other studies used binomial errors with only patients as random effects. Where possible we recalculated the statistical errors on these other studies to include sites or radiologists as

random effects, which resulted in wider estimated confidence intervals. To be conservative we used these wider confidence intervals so that we would not draw conclusions of significance where there were none. The methods for calculating confidence intervals were the same for all FOMs. In our tables we indicate which studies we corrected.

The errors on FPF values were particularly likely to be underestimated by study authors because the sample sizes of non-diseased patients were orders of magnitude larger than the number of readers or sites. Therefore the true errors are dominated by the finite sample sizes of readers or sites, not by the patient sample size. These underestimates of errors were not as significant in the true positive fractions because the diseased sample sizes are much smaller due to the low prevalence of disease. The number of diseased cases is not much greater than the number of readers, so the reader and patient sample sizes are approximately the same, and the binomial assumptions are valid. Likewise the errors for AUC and EU are dominated by the variability in the small diseased patient sample so they are not grossly underestimated.

When calculating errors on AUC and EU estimates that were derived from single reported TPF and FPF values, TPF and FPF were assumed to be independent. Though we expect TPF and FPF to be correlated, correlations measured in other studies were not high, and the independence assumption gave us error bars that should be accurate or conservative. Because every study reported results and errors differently, the specific method used to calculate every estimate or standard error can not be reported here, but it can be obtained from the corresponding author.

The statistical errors on many of the laboratory reader studies are smaller than those on the much larger clinical studies. This is because in laboratory studies extra diseased patients are added, and multiple readers review the same patients retrospectively. This gives many readings of diseased cases, which are relatively rare in observational clinical studies.

2.4. Imaging Studies

In this paper we examined twenty studies that compared reader performance between two breast imaging modalities. We chose these imaging modalities because reports were available for both small laboratory and very large clinical studies. Our methods of selecting studies were not rigorous nor pre-planned. Studies were selected based on internet literature searches and availability of data. No studies that we found were excluded from our sample based on study results.

Nine reader studies compared the performance of readers using screen film mammography (SFM) against their performance with full field digital mammography (FFDM). Table 1 lists these studies and their characteristics. Different studies used different brands of mammography devices. Five of these studies were performed in a laboratory setting, and four were prospective clinical studies. Two of these studies were used to support the initial FDA approval of FFDM devices.

We also reviewed the results of four studies that examined the effect of adding ultrasound imaging to breast cancer screening. These studies compared the performance of X-ray

mammography (XRM) alone to X-ray mammography with ultrasound (XRM+US) in women with dense breasts. These studies and their characteristics are listed in Table 2. Two of these studies were controlled laboratory studies, and two of these studies were clinical trials. The different studies used different brands of imaging devices.

In Table 3 we list eight published studies that compared the performance of radiologists screening patients for breast cancer with two-dimensional digital mammography (DM) images against their performance screening with Hologic (Bedford, Massachusetts, USA) digital breast tomosynthesis images in addition to the DM images (DM+DBT). Three of these studies were controlled enriched laboratory studies. These studies enriched their patient populations with difficult non-diseased patients as well as enriching with breast cancer patients. Five studies examined the performance of DM and DBT in clinical settings.

We specifically indicated European studies in the tables and figures, because a number of differences exist in breast cancer screening between Europe and the United States. These differences include disease prevalence, lesion sizes, recall rates, and the use of double readings.

3. Results

In Figure 2 we plotted the average rTPF estimates for the twenty studies in our sample. The values are plotted horizontally. The error bars indicate the approximate 95% confidence intervals on those mean measures. A rTPF with a value of 1 indicates equivalence between the two modalities within a study. The clinical studies have open circles or boxes as plotting markers. The laboratory reader studies have solid markers. The studies are grouped by the types of modalities that were compared.

Figure 3 plots the mean relative FPF for the twenty studies in our sample. The layout of this plot is the same as Figure 2. Figure 4 plots the mean relative AUC values, and Figure 5 plots the mean rEU values for the same studies. Within each figure and each grouping, we expect to observe consistent values if the changes in FOMs are reproducible across studies.

For a measure of quantitative reproducibility of these relative TPF, FPF, AUC and EU values we calculated the probability of observing the values if they truly all shared a common mean effect size. For each type of study (FFDM/SFM, US/XRM, DBT/DM) within each figure, we performed an approximate χ^2 test. These results are given in Table 4. Small values in the table indicate a low probability that the reported changes in FOMs are truly measuring the same quantity.

In the top eight studies in Figure 2 we see that many of the 95% confidence intervals encompass the value of 1, indicating that the rTPF estimates are consistent with the hypothesis that FFDM and SFM have equal sensitivities. The rTPF confidence intervals are mostly overlapping indicating that the increase (or non-increase) in TPF was reproducible across these FFDM/SFM studies. The probability of observing the rTPF values if they truly all shared a common mean effect size is $p=0.17$, and is given in the second column and row of Table 4. Because this probability is not exceptionally low, we do not have evidence that rTPF is not reproducible in studies comparing FFDM and SFM.

In Figure 2 all the ultrasound studies demonstrate a qualitative increase in TPF with the addition of ultrasound to mammography because all the rTPF values are greater than one. Quantitatively however the amount that TPF increased among studies has a relatively low probability of being the same ($p=0.003$). The χ^2 tests also indicate a very low probability of rTPF having a consistent value across all tomosynthesis studies in the second column of Table 4. The rTPF values do not even agree qualitatively. The retrospective laboratory studies indicate no increase in TPF with the addition of tomosynthesis, whereas the clinical studies strongly indicate an increase, as seen in Figure 2.

The rFPF values in Figure 3 fail to show quantitative reproducibility within any of the study modality groups. All rFPF probabilities in Table 4 are less than 10^{-4} .

Our four ultrasound studies are neither qualitatively nor quantitatively consistent regarding the change in FPF with the addition of ultrasound. The laboratory studies are consistent with no increase in the false positive rate (rFPF=1), but the clinical studies demonstrate large rFPF values. Even though the confidence intervals on the rFPF FOM from Brem et al. [66] may be underestimated, larger confidence intervals would still yield extremely low χ^2 p-values.

Qualitatively across the tomosynthesis studies rFPF is consistently less than one, but quantitatively some studies indicate small reductions in the false positive rate, while other studies indicate larger false positive reductions. The probability that rFPF is quantitatively reproducible across tomosynthesis studies is very small.

rAUC (Figure 4) and rEU (Figure 5) appear to be reproducible within each group of studies. The relative performance of FFDM to SFM is consistent with a value of 1 or a bit less for both rAUC and rEU. All ultrasound studies have an AUC increase consistent with approximately 14% and an increase in EU of approximately 43%. All DBT studies are indicative of an increase in AUC of around 10%, and an increase in EU of 31%. These studies demonstrate that the rEU effect sizes are larger than rAUCs, but the confidence intervals are also larger, so the inferences are frequently similar, as was noted by Abbey et al. [27]. None of the rAUC or rEU χ^2 probabilities in Table 4 are atypically low given the multiple measurements, indicating that rAUC and rEU may be reproducible metrics across many types of imaging studies.

We performed analyses to determine the sensitivity of our results to the assumed prevalences and models. The systematic errors in rAUC due to our uncertainty in disease prevalence are less than the statistical errors in almost all the observational clinical studies. For example in the Rose et al. study [69] doubling an assumed prevalence from 0.0055 to 0.0110 changes the estimated rAUC by 0.015. This change is significantly smaller than the estimated statistical uncertainty on that rAUC, 0.05, and therefore our uncertainty in prevalence can be neglected. The exception was the Brem et al. [66] study where the uncertainty in rAUC due to the uncertainty in prevalence was estimated to be approximately equal to the finite sample error. For this study the rAUC confidence interval reflects this extra uncertainty. We also performed analyses with bi-normal ROC models rather than power-law models. These analyses indicate that our results were not sensitive to the choice of ROC model [40].

As we mentioned previously, the uncertainties on the rFPF values from some of the observational clinical studies may be substantially underestimated because variation among readers or study sites was not considered and we could not correct them. Such underestimated uncertainties inflate the χ^2 test values, and depress the p-values. In an extra analysis we inflated the uncertainties on these studies. This inflation did not increase p-values significantly except for the rFPF value for the FFDM/SFM studies, which may be greater than 10^{-4} , almost 0.007.

4. Discussion

While there has been much discussion in the literature about which figures of merit are most powerful or clinically relevant for reader studies of diagnostic imaging devices, a fundamental requirement of any FOM is its ability to measure differences between experimental conditions reproducibly. If a change in an FOM can not be reproduced across multiple scientific studies by different researchers, or if a measured change of an FOM in a laboratory study is not the same as a later clinical study, then there is little reason to report it in pre-clinical studies.

Across different types of reader imaging studies we expect that rTPF or rFPF may not be reproducible because these metrics depend on decision thresholds which depend upon prevalence and costs. For example when discussing meta-analyses of diagnostic studies Irwig et al. [74] stated,

If primary studies provide only sufficient information to estimate sensitivity and specificity, the mean sensitivity and the mean specificity can be estimated, possibly weighted in some way for the sample size of each study. However, this technique is inappropriate because it is likely that different studies use different explicit or implicit thresholds, so that a primary study with a high sensitivity may have a low specificity and vice versa.

Indeed across the different studies we made no effort to compensate for explicit or implicit thresholds when measuring changes in TPF or FPF. We used the values reported by the studies.

Most studies provide a control arm against which we can compare the experimental modality using the same readers, reading environment, and patient population. Therefore we might expect that a *change* in reported TPF or FPF between those two arms may be reproducible across different types of reader imaging studies. However, our meta-analyses demonstrate that this is not true. The changes in TPF and FPF between study arms are frequently significantly different across studies that examine the same or similar pairs of modalities both qualitatively and quantitatively. The significance of these differences remains even if we account for the twelve multiple comparisons in Table 4 using an overly conservative Bonferroni correction [75].

Gur [76] noted that equal changes in TPF in two different studies does not necessitate an equal change in diagnostic accuracy (AUC) between the two studies. This change is dependent upon the readers' operating points on their ROC curve. The same applied to

changes in FPF. Conversely we have shown that equal changes in diagnostic accuracy (AUC) or utility between studies does not imply equal changes in TPF or FPF.

AUC does not depend on a decision threshold, or changes in a threshold between modalities within a study. Therefore we found that the relative increase in AUC, rAUC, is reproducible across laboratory and clinical studies for several comparisons of imaging modalities. AUC can be considered a measure of diagnostic information [77] that the imaging device gives the reader. Even if devices are not used the same way in a controlled laboratory studies and clinical observational studies, rAUC is a measure of the increase in information that readers obtained using one diagnostic over another, and that is the same regardless of the study design.

EU is a weighted function that trades TPF with FPF at a particular decision threshold. This trade-off and fixed threshold make EU a stable and reproducible measure across many study types.

Our ability to infer changes of diagnostic performance reproducibly across laboratory and observational clinical studies may influence the types of studies we decide to perform. For example, laboratory studies should always plan to report a FOM that is known to be reproducible, if we want their results to hold in later clinical practice. If there is a standard imaging diagnostic, and a pre-clinical laboratory study of a new diagnostic demonstrates an increased AUC or EU, then we know that the new device would improve that diagnostic performance in a clinical setting also. This reduces the need for large clinical studies, or alternatively the need for large clinical studies that follow diagnostically negative patients.

5. Conclusions

To summarize we have demonstrated that we can make estimates of relative sensitivity (TPF), false positive rate, AUC, and expected utility between two arms of a imaging reader study, even if that study is an observational clinical study that does not follow negatively called patients. We can compare those estimates across many types of imaging studies with different characteristics, including small, retrospective, laboratory, reader studies and large, observational, clinical studies.

Our comparisons across twenty studies demonstrate that observed changes in sensitivity (TPF), specificity, or the false positive fraction (FPF) between study arms are not reproducible either qualitatively or quantitatively. However changes in AUC or expected utility appear to be very reproducible.

Acknowledgments

This work was supported in part by the National Institutes of Health Grant R21 EB018939.

The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

References

1. Swets, JA., Pickett, RM. Evaluation of diagnostic systems: methods from signal detection theory. Academic Press; New York: 1982.
2. Metz CE, Wagner RF, Doi K, Brown DG, Nishikawa RM, Myers KJ. Toward consensus on quantitative assessment of medical imaging systems. *Medical Physics*. 1995; 22:1057–61. [PubMed: 7565380]
3. Gur D, Rockette HE, Warfel T, Lacomis JM, Fuhrman CR. From the laboratory to the clinic: The prevalence effect. *Academic Radiology*. 2003; 10:1324–1326. [PubMed: 14626308]
4. Metz CE. ROC analysis in medical imaging: a tutorial review of the literature. *Radiol Phys Technol*. 2008; 1(1):2–12. [PubMed: 20821157]
5. Gallas BD, Chan HP, DOrsi CJ, Dodd LE, Giger ML, Gur D, Krupinski EA, Metz CE, Myers KJ, Obuchowski NA, Sahiner B, Toledano AY, Zuley ML. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Academic Radiology*. 2012; 19:463–477. [PubMed: 22306064]
6. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology*. 2003; 229(1):3–8. [PubMed: 14519861]
7. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev Vet Med*. 2000; 45(1-2):23–41. [PubMed: 10802332]
8. Begg CB. Experimental design of medical imaging trials. Issues and options. *Invest Radiol*. 1989; 24(11):934–936. [PubMed: 2807813]
9. Pinsky PF, Gallas B. Enriched designs for assessing discriminatory performance—analysis of bias and variance. *Stat Med*. 2012; 31(6):501–515. [PubMed: 22095795]
10. Hilden J. The area under the ROC curve and its competitors. *Med Decis Making*. 1991; 11(2):95–101. [PubMed: 1865785]
11. Moons KG, Stijnen T, Michel BC, Buller HR, Van Es GA, Grobbee DE, Habbema JD. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. *Med Decis Making*. 1997; 17(4):447–454. [PubMed: 9343803]
12. Hilden J. Evaluation of diagnostic tests - the schism. *Society for Medical Decision Making Newsletter*. 2004:5–6.
13. Gur D, Bandos AI, Rockette HE, Zuley ML, Hakim CM, Chough DM, Ganott MA, Sumkin JH. Is an ROC-type response truly always better than a binary response in observer performance studies? *Academic Radiology*. 2010; 17(5):639–645. [PubMed: 20236840]
14. Hand DJ, Anagnostopoulos C. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*. 2013; 34:492–495. DOI: 10.1016/j.patrec.2012.12.004
15. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *Eur Radiol*. 2015; 25(4):932–939. [PubMed: 25599932]
16. Skaane P, Niklason L. Receiver operating characteristic analysis: A proper measurement for performance in breast cancer screening? *Am J Roentgenol*. 2006; 354:579–580. DOI: 10.2214/AJR.06.5007
17. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; Oxford: 2003.
18. Zhou, X., McClish, D., Obuchowski, N. *Wiley Series in Probability and Statistics*. Wiley; 2009. *Statistical Methods in Diagnostic Medicine*.
19. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987; 6(4):411–423. [PubMed: 3114858]
20. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*. 1978; 7(4):283–298.
21. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]

22. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol.* 1986; 21:720–733. [PubMed: 3095258]
23. Patton DD, Woolfenden JM. A utility-based model for comparing the cost-effectiveness of diagnostic studies. *Invest Radiol.* 1989; 24(4):263–271. [PubMed: 2501232]
24. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology.* 2005; 16(1):73–81. [PubMed: 15613948]
25. Halpern EJ, Albert M, Krieger AM, Metz CE, Maidment AD. Comparison of receiver operating characteristic curves on the basis of optimal operating points. *Acad Radiol.* 1996; 3(3):245–253. [PubMed: 8796672]
26. Abbey CK, Eckstein MP, Boone JM. Estimating the relative utility of screening mammography. *Medical Decision Making.* 33
27. Abbey CK, Gallas BD, Boone J, Niklason LT, Hadjiiski LM, Sahiner B, Samuelson FW. Comparative statistical properties of expected utility and area under the ROC curve for laboratory studies of observer performance. *Academic Radiology.* 2014; 21:481–490. [PubMed: 24594418]
28. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology.* 1996; 201(3):745–750. [PubMed: 8939225]
29. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. *AJR Am J Roentgenol.* 2001; 177(3):543–549. [PubMed: 11517044]
30. Ma H, Bandos AI, Rockette HE, Gur D. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat Med.* 2013; 32(20):3449–3458. [PubMed: 23508757]
31. Keilwagen J, Grosse I, Grau J. Area under precision-recall curves for weighted and unweighted data. *PLoS ONE.* 2014; 9(3):e92209. [PubMed: 24651729]
32. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A free-response approach to the measurement and characterization of radiographic observer performance. *Journal of Applied Photographic Engineering.* 1978; 3:166–171.
33. Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Medical Physics.* 1989; 16:561–568. [PubMed: 2770630]
34. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys.* 1996; 23(10):1709–1725. [PubMed: 8946368]
35. Zheng B, Chakraborty DP, Rockette HE, Maitz GS, Gur D. A comparison of two data analyses from two observer performance studies using Jackknife ROC and JAFROC. *Med Phys.* 2005; 32(4):1031–1034. [PubMed: 15895587]
36. Chakraborty DP. Validation and statistical power comparison of methods for analyzing free-response observer performance studies. *Acad Radiol.* 2008; 15(12):1554–1566. [PubMed: 19000872]
37. Popescu LM. Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve. *Med Phys.* 2011; 38(10):5690–5702. [PubMed: 21992384]
38. C. o. A. National Academies of Sciences. *T Statistics, Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop.* National Academies Press; 2016.
39. Schatzkin A, Connor RJ, Taylor PR, Bunnag B. Comparing new and old screening tests when a confirmatory procedure cannot be performed on all screens. *American Journal of Epidemiology.* 1987; 125(4):672–678. [PubMed: 3826045]
40. Samuelson FW, Abbey CK. Using relative statistics and approximate disease prevalence to compare screening tests. *International Journal of Biostatistics.*
41. Abbey CK, Eckstein MP, Boone JM. An equivalent relative utility metric for evaluating screening mammography. *Medical Decision Making.* 2010; 30:113–122. [PubMed: 19706880]
42. Tanner WP, Swets JA. A decision-making theory of visual detection. *Psychological Review.* 1954; 61:401–409. [PubMed: 13215690]
43. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in a laboratory environment: Changing the confidence ratings. *Academic Radiology.* 2007; 14:49–53. [PubMed: 17178365]

44. Samuelson FW. Inference based on diagnostic measures from studies of new imaging devices. *Academic Radiology*. 2013; 20:816–824. [PubMed: 23643364]
45. Rafferty EA, Park JM, Philpotts LE, Poplack SP, Sumkin JH, Halpern EF, Niklason LT. Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: Results of a multicenter, multireader trial. *Radiology*. 2013; 266:104–113. DOI: 10.1148/radiol.12120674 [PubMed: 23169790]
46. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. *Journal of Mathematical Psychology*. 1969; 6:487.
47. Metz, CE. Statistical analysis of ROC data in evaluating diagnostic performance. In: Herbert, D., Myers, R., editors. *Multiple regression analysis: applications in the health sciences*. American Institute of Physics; New York: 1986. p. 365
48. Egan, JP. *Signal Detection Theory and ROC analysis*. Academic Press; New York: 1975.
49. Samuelson FW, He X. A comparison of semi-parametric ROC models on observer data. *SPIE Journal of Medical Imaging*. 2014; 1(3):031004.doi: 10.1117/1.JMI.1.3.031004 [PubMed: 26158046]
50. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, D'Orsi C, Jong R, Rebner M. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*. 2005; 353:1773–1783. [PubMed: 16169887]
51. Rosenberg RD, Yankaskas BC, Abraham LA, Sickles EA, Lehman CD, Geller BM, Carney PA, Kerlikowske K, Buist DSM, Weaver DL, Barlow WE, Ballard-Barbash R. Performance benchmarks for screening mammography. *Radiology*. 2006; 241(1):55–66. [PubMed: 16990671]
52. Dorfman R. A note on the δ -method for finding variance formulae. *The Biometric Bulletin*. 1938; 1:129–137.
53. Crmer, H. *Mathematical Methods of Statistics*. Princeton University Press; Princeton, NJ: 1946.
54. Kerlikowske K, Hubbard RA, Miglioretti DL, Geller BM, Yankaskas BC, Lehman CD, Taplin SH, Sickles EA. Comparative effectiveness of digital versus film-screen mammography in community practice in the united states: A cohort study. *Annals of Internal Medicine*. 2011; 155:493–502. DOI: 10.7326/0003-4819-155-8-201110180-00005 [PubMed: 22007043]
55. Cole E, Pisano ED, Brown M, Kuzmiak C, Braeuning MP, Kim HH, Jong R, Walsh R. Diagnostic accuracy of Fischer Senoscan Digital Mammography versus screen-film mammography in a diagnostic mammography population. *Acad Radiol*. 2004; 11(8):879–886. [PubMed: 15288038]
56. F. I. Corporation. Tech rep. Fischer Imaging Corporation; 2001. Summary of safety and effectiveness data, P010017. URL http://www.accessdata.fda.gov/cdrh_docs/pdf/P010017b.pdf
57. Hendrick, RE., Lewin, JM., D'Orsi, CJ., Kopans, DM., Conant, E., Cutter, GR., Sitzler, A. Non-inferiority study of FFDM in an enriched diagnostic cohort: comparison with screen-film mammography in 625 women. In: Yaffe, MJ., editor. *5th International Workshop on Digital Mammography*. 2000. p. 475–481.
58. Systems, GM. Tech rep. GE Medical Systems; 2000. Summary of safety and effectiveness data, P990066. URL http://www.accessdata.fda.gov/cdrh_docs/pdf/P990066B.pdf
59. Hendrick RE, Cole EB, Pisano ED, Acharyya S, Marques H, Cohen MA, Jong RA, Mawdsley GE, Kanal KM, D'Orsi CJ, Rebner M, Gatsonis C. Accuracy of soft-copy digital mammography versus that of screen-film mammography according to digital manufacturer: ACRIN DMIST retrospective multireader study. *Radiology*. 2008; 247(1):38–48. [PubMed: 18372463]
60. Lewin JM, D'Orsi CJ, Hendrick RE, Moss LJ, Isaacs PK, Karellas A, Cutter GR. Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. *AJR Am J Roentgenol*. 2002; 179(3):671–677. [PubMed: 12185042]
61. Skaane P, Hofvind S, Skjennald A. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: follow-up and final results of Oslo II study. *Radiology*. 2007; 244(3):708–717. [PubMed: 17709826]
62. Giger ML, Inciardi MF, Edwards A, Papaioannou J, Drukker K, Jiang Y, Brem R, Brown JB. Automated breast ultrasound in breast cancer screening of women with dense breasts: Reader

- study of mammography-negative and mammography-positive cancers. *AJR Am J Roentgenol*. 2016; 206(6):1341–1350. [PubMed: 27043979]
63. U-Systems Inc. Tech rep. U-Systems Inc; 2012. Summary of safety and effectiveness data, P110006. URL http://www.accessdata.fda.gov/cdrh_docs/pdf11/P110006b.pdf
 64. Kelly KM, Dean J, Lee SJ, Comulada WS. Breast cancer detection: radiologists' performance using mammography with and without automated whole-breast ultrasound. *Eur Radiol*. 2010; 20(11): 2557–2564. [PubMed: 20632009]
 65. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-Velez M, Pisano ED, Jong RA, Evans WP, Morton MJ, Mahoney MC, Larsen LH, Barr RG, Farria DM, Marques HS, Boparai K. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *Journal of the American Medical Association*. 2008; 299(18):2151–2163. [PubMed: 18477782]
 66. Brem RF, Tabár L, Duffy SW, Inciardi MF, Guingrich JA, Hashimoto BE, Lander MR, Lapidus RL, Peterson MK, Rapelyea JA, Roux S, Schilling KJ, Shah BA, Torrente J, Wynn RT, Miller DP. Assessing improvement in detection of breast cancer with three-dimensional automated breast ultrasound in women with dense breast tissue: The SonoInsight study. *Radiologist*. 2015; 274(3): 663–673.
 67. Gur D, Abrams GS, Chough DM, Ganott MA, Hakim CM, Perrin RL, Rathfon GY, Sumkin JH, Zuley ML, Bandos AI. Digital breast tomosynthesis: Observer performance study. *American Journal of Roentgenology*. 2009; 193:586–591.
 68. Inc, H. Tech rep. Hologic Inc; 2011. Summary of safety and effectiveness data, P080003. URL http://www.accessdata.fda.gov/cdrh_docs/pdf8/P080003B.pdf
 69. Rose SL, Tidwell AL, Bujnoch LJ, Kushwaha AC, Nordmann AS, Sexton R. Implementation of breast tomosynthesis in a routine screening practice: An observational study. *American Journal of Roentgenology*. 2013; 200:1401–1408. DOI: 10.2214/AJR.12.9672 [PubMed: 23701081]
 70. Friedewald SM, Rafferty EA, Rose SL, Durand MA, Plecha DM, Greenberg JS, Hayes MK, Copit DS, Carlson KL, Cink TM, Barke LD, Greer LN, Miller DP, Conant EF. Breast cancer screening using tomosynthesis in combination with digital mammography. *Journal of the American Medical Association*. 2014; 311(24):2499–2507. [PubMed: 25058084]
 71. Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, Izadi M, Jebsen IN, Jahr G, Krager M, Niklason LT, Hofvind S, Gur D. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*. 2013; 267(1):47–56. [PubMed: 23297332]
 72. Ciatto S, Houssami N, Bernardi D, Caumo F, Pellegrini M, Brunelli S, Tuttobene P, Bricolo P, Fanto C, Valentini M, Montemezzi S, Macaskill P. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. *Lancet Oncol*. 2013; 14(7):583–589. [PubMed: 23623721]
 73. Haas BM, Kalra V, Geisel J, Raghu M, Durand M, Philpotts LE. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. *Radiology*. 2013; 269(3):694–700. [PubMed: 23901124]
 74. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994; 120(8):667–676. [PubMed: 8135452]
 75. Dunn OJ. Multiple comparisons among means. *Journal of the American Statistical Association*. 1961; 56(293):52–74. DOI: 10.1080/01621459.1961.10482090
 76. Gur D. Imaging technology and practice assessment studies: Importance of the baseline or reference performance level. *Radiology*. 2008; 247:8–11. DOI: 10.1148/radiol.2471070822 [PubMed: 18372460]
 77. Shen F, Clarkson E. Using Fisher information to approximate ideal observer performance on detection tasks for lumpy-background images. *Journal of the Optical Society of America A*. 2006; 23(10):2406–2414.

Figure 1.

The solid line is an ROC curve, giving the trade-off between the true and false positive fractions as a decision threshold is varied on the output of a diagnostic. The shaded area below the curve is the AUC. The length of dotted vertical line is the expected utility (EU). It is the intercept of the dashed line of constant utility that is tangent to the ROC curve at the optimal operating point (circle) [26].

Figure 2.

The relative true positive fraction, rTPF, estimated from eight studies comparing full-field digital mammography and screen-film mammography, four studies comparing X-ray mammography with and without ultrasound, and eight studies comparing digital mammography with and without Hologic digital breast tomosynthesis imaging. These studies are listed in Tables 1, 2, and 3. Open circles or squares indicate observational clinical studies. Squares are European studies. Solid circles indicate controlled laboratory reader studies. Equal performance of the two modalities in each study is indicated by the vertical line at 1.0. Horizontal bars are approximate 95% confidence intervals on the mean value.

Figure 3.

The relative false positive fraction, rFPF, estimated from the 20 studies listed in Tables 1, 2, and 3. See the caption of Figure 2 for other details. Error bars on some of the clinical studies are probably underestimated.

Figure 4.

The relative area under the ROC curve, rAUC, estimated from the 20 studies listed in Tables 1, 2, and 3. The Berg et al. [65] study calculated AUC in two ways, and both are shown. See the caption of Figure 2 for other details.

Figure 5.

The relative expected utility, rEU, estimated from the 20 studies listed in Tables 1, 2, and 3. See the caption of Figure 2 for other details.

Table 1

Studies that compared full-field digital mammography (FFDM) against screen film mammography (SFM). “Lab” indicates a retrospective laboratory reader study; “Clinic” indicates a prospective clinical study. “FDA app” indicates that data from the study were part of a FDA device approval application. “Euro” indicates European studies. In the analysis column, “A” indicates studies that did not include sites or readers as random effects. “B” indicates the same, but where we recalculated the errors. “C” indicates studies that followed patients or used retrospective case selection, so disease prevalence was known. “D” indicates that studies reported AUC values.

Study References	Notes	Type of study	Number of patients	Prevalence	Number of readers	Notes on analyses
Cole et al. [55, 56]	Fischer FDA app.	lab	247	0.45	8	BCD
Hendrick et al. [57, 58]	GE FDA app.	lab	625	0.07	5	CD
Hendrick et al. [59, 27]	Fischer FFDM	lab	115	0.37	6	CD
Hendrick et al. [59, 27]	Fuji FFDM	lab	98	0.28	12	CD
Hendrick et al. [59, 27]	GE FFDM	lab	120	0.40	12	CD
Pisano et al. [50]	DMIST	clinic	42,760	0.006	160	ACD
Lewin et al. [60]		clinic	6,736	0.006		ACD
Skaane et al. [61]	Oslo II, Euro	clinic	23,929	0.005	8	AC
Kerlikowske et al. [54]	BCSC	clinic	329,261	0.005	~800	C

Table 2

Studies that compared X-ray mammography with breast ultrasound (XRM+US) against X-ray mammography alone (XRM) on women with dense breasts. Prevalence values in parentheses were assumed values, general screening population of dense-breasted women. See the caption of Table 1 for other details.

Study References	Notes	Type of study	Number of patients	Prevalence	Number of readers	Notes on analyses
Giger et al. [62, 63]	ABUS FDA app.	lab	185	0.28	17	CD
Kelly et al. [64]	AWBU	lab	102	0.50	12	CD
Berg et al. [65]	ACRIN 6666	clinic	2,637	0.015 (0.01)	>21	CD
Brem et al. [66]	SomoInsight	clinic	15,318		39	A

Table 3

Studies that compared digital mammography and Hologic breast tomosynthesis (DM+DBT) against digital mammography alone (DM). See the caption of Table 1 for other details. Prevalence values in parentheses were assumed values.

Study References	Notes	Type of study	Number of patients	Prevalence	Number of readers	Notes on analyses
Gur et al. [67, 13, 68]	FDA app.	lab	125	0.28	8	CD
Rafferty et al. [45, 68, 27]	FDA app.	lab	312	0.15	12	CD
Rafferty et al. [45, 68, 27]	FDA app.	lab	312	0.16	15	CD
Rose et al. [69]		clinic	23,355	(0.0064)	6	B
Friedewald et al. [70]		clinic	454,850	(0.0064)	139	B
Skaane et al. [71]	OTST, Euro	clinic	12,621	0.0096	8	BC
Ciatto et al. [72]	STORM, Euro	clinic	7,292	(0.0095)	8	A
Haas et al. [73]		clinic	13,158	(0.0095)	8	A

Table 4

Probabilities that the mean measured relative statistic from all studies of a modality are consistent with a common mean value based on an approximate χ^2 test. These probabilities were calculated for each figure of merit and modality, and no corrections for multiple comparisons were performed. The starred value may be underestimated.

	rTPF	rFPF	rAUC	rEU
FFDM/SFM studies	0.17	<10 ⁻⁴ *	0.04	0.09
US/XRM studies	0.003	<10 ⁻⁴	0.39	0.09
DBT/DM studies	<10 ⁻⁴	<10 ⁻⁴	0.87	0.77