Full Title Page

Title: "Impact of Prevalence and Case Distribution in Lab-based Diagnostic Imaging Studies"

This work is an original investigation article.

Authors and Affiliations

- Brandon D. Gallas, PhD [a]
  - **Corresponding Author**
  - brandon.gallas@fda.hhs.gov
  - 301-796-2531 (office)
  - 301-796-9925 (fax)
  - 10903 New Hampshire Ave, WO62-4104, Silver Spring, MD, 20993
- Weijie Chen, PhD [a]
  - Weijie.Chen@fda.hhs.gov
- Elodia Cole, MS [b]
  - ebcole.be@gmail.com
- Robert Ochs, PhD [c]
  - Robert.Ochs@fda.hhs.gov
- Nicholas A. Petrick, PhD [a]
  - Nicholas.Petrick@fda.hhs.gov
- Etta D. Pisano, MD [b]
  - etpisano@gmail.com
- Berkman Sahiner, PhD [a]
  - Berkman.Sahiner@fda.hhs.gov
- Frank W. Samuelson, PhD [a]
  - Frank.Samuelson@fda.hhs.gov
- Kyle J. Myers, PhD [a]
  - Kyle.Myers@fda.hhs.gov

Affiliations:

a. **Institution from which the work originated**: FDA/CDRH/OSEL/Division of Imaging, Diagnostics, and Software Reliability, 10903 New Hampshire Ave, WO62-4104, Silver Spring, MD, 20993-0002.
b. Medical University of South Carolina, 171 Ashley Ave, Charleston, SC 29425.
c. FDA/CDRH/OIR/Division of Radiological Health, 10903 New Hampshire Ave, WO66-4312, Silver Spring, MD, 20993-0002.

Abstract

Ratonale and Objectives We investigated effects of prevalence and case distribution on radiologist diagnostic performance as measured by area under the receiver operating characteristic curve (AUC) and sensitivity-specificity in lab-based reader studies evaluating imaging devices.

Materials and Methods Our retrospective reader studies compared full-field digital mammography (FFDM) to screen-film mammography (SFM) for women with dense breasts. Mammograms were acquired from the prospective Digital Mammographic Imaging Screening Trial (DMIST). We performed five reader studies that differed in terms of cancer prevalence and the distribution of non-cancers. Twenty radiologists participated in each reader study. Using split-plot study designs, we collected recall decisions and multi-level scores from the radiologists for calculating sensitivity, specificity, and AUC.

Results Differences in reader-averaged AUCs slightly favored SFM over FFDM (biggest AUC difference: 0.047, SE=0.023 p=0.047), where standard error (SE) accounts for reader and case variability. The differences were not significant at a level of 0.01 (0.05/5 reader studies). The differences in sensitivities and specificities were also indeterminate. Prevalence had little effect on AUC (largest difference: 0.02), whereas sensitivity increased and specificity decreased as prevalence increased.

Conclusion We found that AUC is robust to changes in prevalence, while radiologists were more aggressive with recall decisions as prevalence increased.

Keywords Image Evaluation, Study Design, MRMC analysis, AUC, Sensitivity, Specificity

1. Introduction

Lab-based reader studies are often used to evaluate imaging technologies and are typically characterized by 1) a moderate number of cases that may not represent the true patient population (e.g., higher disease prevalence because of enrichment); 2) structured and quantitative case reports based on a narrow task that is often a simplification of the clinical task (e.g., the radiologist only evaluates the images and is blinded to patient information); and 3) retrospective reads with no impact on patient management. Here we report on lab-based reader studies in a project we refer to as VIPER, Validation of Imaging Premarket Evaluation and Regulation. The setting for this validation is the comparison of a new imaging modality to a reference imaging modality. The data from such studies may be used to support Food and Drug Administration (FDA) clearance or approval of medical imaging devices and computer aids. VIPER was born from a desire to validate the use of lab-based studies as an alternative to large prospective clinical trials.

One of the largest prospective clinical imaging trials with data available conducted at the time of VIPER's conception (2010) was the Digital Mammographic Imaging Screening Trial (DMIST) (1,2). DMIST was designed to compare full-field digital mammography (FFDM) to screen-film mammography (SFM), pooling results from five different FFDM platforms and six different SFM platforms. DMIST was sized to detect an AUC difference of 0.06 between FFDM and SFM with 5% Type I error and 80% power. This requirement and the low prevalence of breast cancer in the screening population drove the study to enroll 49,528 women, with all relevant information obtained for 42,760 subjects. Such a study is too expensive for a single manufacturer or investigator to afford, so many turn to lab-based reader studies.

For decades psychophysical signal-detection experiments have demonstrated the effects of prevalence on decision threshold (3,4). Egglin and Feinstein (5) found significant differences in sensitivity, specificity, and AUC of radiologists diagnosing pulmonary emboli at 20% and 60% prevalence. Gur et al (6,7) compared radiologists' performance in diagnosing several different pathologies in several different prevalence conditions and found no significant differences in AUC but found significant shifts in those radiologists' ratings (8). However, Gur et al. (9) did find a significant difference in performance between mammographers interpreting screening mammograms in the clinic and in a laboratory study with higher prevalence. Evans et al. (10) demonstrated the effects of cytologists' decision thresholds as a function of prevalence of cervical cancers. With different coauthors Evans also showed that radiologists mark more cases as cancer in a highly enriched setting (50% prevalence) compared to when the same cases were inserted into their normal screening service workflow (11). The current work builds on this literature.

The purpose of VIPER is to investigate the effects of prevalence and case-distribution on radiologist performance detecting cancer as measured by AUC and sensitivity-specificity in lab-based reader studies evaluating imaging devices. We use a data-collection method that enforces consistency between the clinical (binary) recall decision and ROC scores, and we compare the sensitivity-specificity operating points measured in VIPER to those found in the prospective DMIST study. Initial results of this paper were presented at the SPIE Medical Imaging Conference (12). The patient population in VIPER is limited to women with dense breasts, which was a DMIST sub-population where FFDM was found to significantly outperform SFM (AUC for FFDM was 0.78, SFM was 0.68). The main VIPER hypotheses test the differences in AUCs from FFDM and SFM across five reader studies. The studies differ in terms of their study populations, namely, the prevalence and the distribution of non-cancer cases.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

2. Materials and Methods

VIPER was conducted at the Medical University of South Carolina (IRB 13890) from September 2013 to August 2015 as a retrospective image evaluation study. The images used in VIPER are from the ACRIN (American College of Radiology Imaging Network) DMIST trial (1,2) and were acquired from ACRIN. In addition to images, ACRIN provided Breast Imaging Reporting and Data System (BIRADS) management

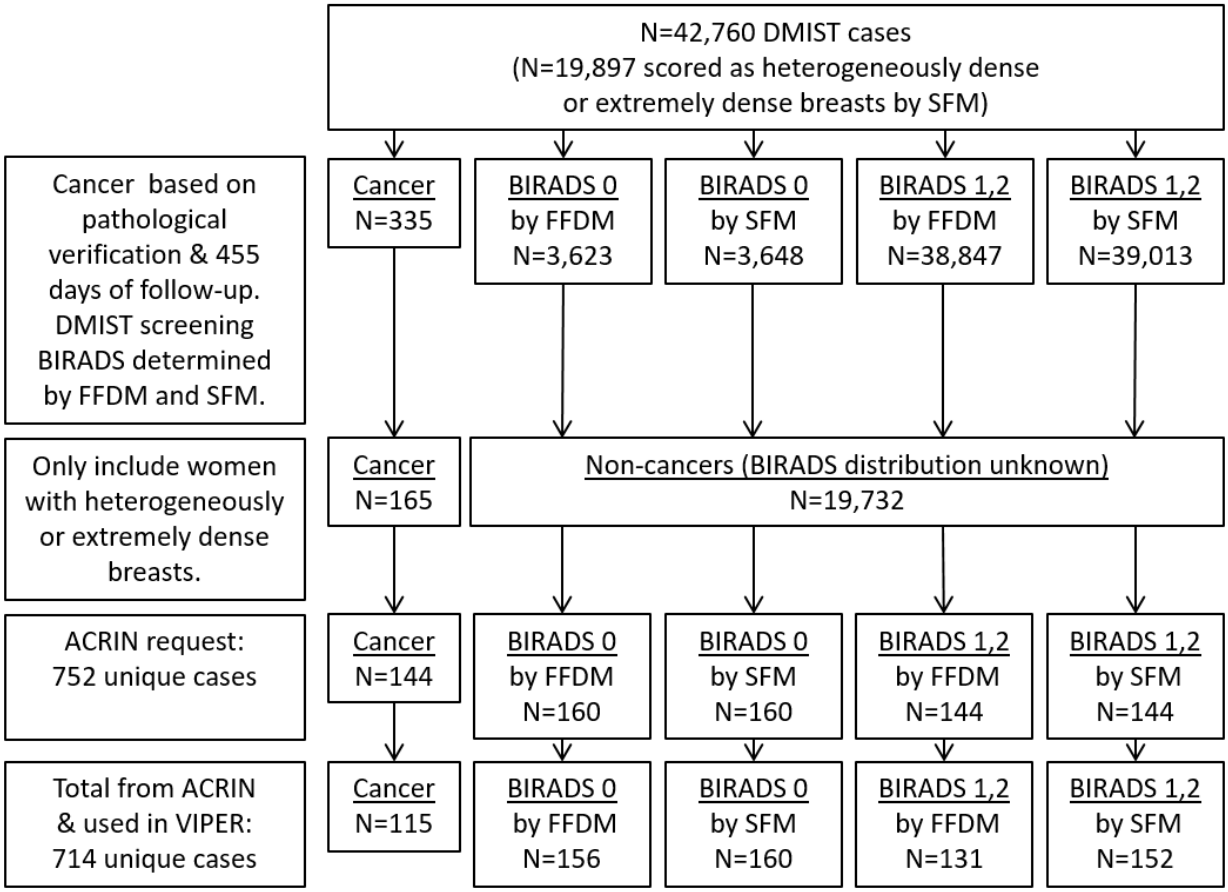

Figure 1: Figure 1: Participant flow diagram. Screening and Challenge sub-studies of different prevalences are created from the final subgroups based on DMIST screening BIRADS scores by FFDM and SFM. The last row differs from the one above it due to availability from ACRIN.

scores and cancer determinations (based on pathological verification for cancer and up to 455 days of follow-up for non-cancer) as per the original DMIST study design. ACRIN removed patient identifiers from all clinical images and data, in accordance with the primary requirements of HIPAA.

2.1 Subjects=Cases

Eligibility for inclusion in VIPER required that the BIRADS breast density classification documented in DMIST be 3 or 4 by SFM (13). These are women with heterogeneously and extremely dense breasts, categories "c" and "d" in the current lexicon (14). The participant flow diagram is given in Figure 1. The age range of the women in VIPER was 33 to 86 (mean: 54).

2.2 Split-Plot Study Design

Given the available cases, we designed five split-plot reader studies (15). In a split-plot study, readers and cases are split into groups and each reader group is assigned a case group. Each split-plot group is, in effect, a fully-crossed study (within that group): a study where every reader reads every case in both modalities. In other words, every reader in reader-group 1 reads every case in case-group 1, every reader in reader-group 2 reads every case in case-group 2, and so on. Each of the VIPER reader studies had four split-plot groups. A

split-plot study can be more efficient statistically and more efficient in terms of resources compared to a fully crossed study (15,16). We describe the reader studies below and provide more details in the Supplementary Materials (17)

2.3 Five VIPER Study Populations

Prevalence: The VIPER reader studies investigated study populations with different prevalences. Table 1 shows the per reader average number of cancers, non-cancers, and cancer prevalences of the five VIPER reader studies. The lowest prevalence in the VIPER reader studies was `10.6`%. For comparison, the prevalence in DMIST (and the subpopulation of women with dense breasts) was much lower, `0.8`%.

Distribution of non-cancer cases: The VIPER reader studies investigated two kinds of non-cancer study populations: a 'screening' population and a 'challenge' population. The non-cancer cases in the screening populations were heavily weighted with BIRADS 1,2 cases, while the non-cancer cases in the challenge populations included only BIRADS 0 cases. BIRADS 0 cases were challenging because, while some were the result of bad image quality, most were thought to be suspicious enough for cancer to request additional evaluations. Table 2 shows the per-reader distributions of BIRADS patient management scores for women without cancer. The ratio in DMIST (10.69) was moderately larger than the ratios in VIPER screening reader studies (5.84 to 8.36).

Table 1 title: Table of average per-reader cancer prevalence.

VIPER Reader Study	Average # of Cancers	Average # of Non-Cancers	Average Prevalence (%)	Total Obs
screeningLowP	18.2	154.4	10.6	6911
screeningMedP	28.6	79.2	26.6	4325
screeningHighP	27.2	32.5	45.6	2390
challengeMedP	28.5	80.9	26.1	4377
challengeHighP	26.9	32.5	45.2	2379

Table 1 footnote: Cancer cases are those that were pathologically verified within 455 days after the initial study mammogram. The average number of cancers and non-cancers are presented here as the case subgroups were not all equal. An "observation" is one radiologist's score for one case. VIPER reader studies only included women with heterogeneously dense or extremely dense breasts. For reference, pooling over all DMIST radiologists, DMIST found 165 cancers in 19,897 women with dense breasts (prevalence: 0.8%).

Table 2 title: Table of per-reader distributions of BIRADS patient management scores for women without cancer.

VIPER Reader Study	BIRADS 1&2	BIRADS 0	ratio
screeningLowP	137.95	16.50	8.36
screeningMedP	70.30	8.95	7.85
screeningHighP	27.75	4.75	5.84
challengeMedP	2.25	78.65	0.03
challengeHighP	2.00	30.55	0.07

Table 2 footnote: Average counts are per reader and based on DMIST SFM and FFDM evaluations. Average counts are presented as the case subgroups were not all equal. For reference, DMIST SFM screening yielded 39,013 BIRADS 1 & 2 women compared to 3648 BIRADS 0 women (ratio: 10.69).

2.4 Radiologists=Readers

There were 20 readers in each VIPER sub-study. Readers were allowed to participate in more than one reader study as long as they were assigned to groups with no overlap in cases across the reader studies. Ultimately, 43 readers participated across the five reader studies.

All readers were American Board of Radiology certified, MQSA qualified, and had clinically interpreted at least 50 FFDM images and 50 SFM images as part of their residency or practice. Based on a reader qualification survey, the median number of years interpreting mammograms post residency was 9 (range: 0 to 30). Additionally, readers tended to read with FFDM more than with SFM in their clinical practice. The number of cases read by all the VIPER readers in the last two years before the study was as follows:

- **Mean:** FFDM = 8892, SFM = 450
- **Median:** FFDM = 8000, SFM = 100
- **Minimum:** FFDM = 990, SFM = 0
- **Maximum:** FFDM = 20,000, SFM = 9000

All the readers traveled to a central reading location at least twice to participate in two reading sessions for each reader study. The minimum washout time between two sessions was 27 days and the median was 49.5. In the first reading session, they read half the cases in FFDM and half in SFM. In the second reading session, each reader independently read the opposite modalities for each of the cases. The case order was randomized within a modality and session, and the modality order was assigned in a balanced way.

2.5 Data Collection

Readers were blinded to patient demographics, patient history, the DMIST BIRADS screening scores, the cancer status, and the other radiologist evaluations. For each of the five VIPER study populations, the prevalence and distribution of non-cancer cases (Tables 1 and 2) were described to the participating radiologists, but case counts in specific categories were not provided. Radiologists evaluated the current screening images only (CC and MLO for both breasts), which is different from DMIST, where priors were available to radiologists.

2.5.1 SFM Images

The original SFM images submitted to ACRIN for the DMIST study were used in this study. The films were between 10-12 years old by the time this reader study started. Films were cleaned of any marks in preparing for the study. Films were presented on the top and bottom rollers of a RadX Mammoscope. Shutters were available to mask extraneous light around smaller 8x10 mammo film sizes. Magnifying glass and hot lamp were available at the lightbox.

2.5.2 FFDM Images

The original image processing algorithms applied during DMIST were not available for this study. We did apply image processing to the images to make them interpretable for this reader study choosing a commercial third party vendor: the algorithm was Adar2D by Real Time Tomography, LLC (www.realtimetomography.com).

All mammograms were displayed on dual Eizo 5MP grayscale monitors following a default hanging protocol. Unfortunately, the monitors were not calibrated for this study. Readers were trained to zoom the image (magnification), pan, adjust brightness and contrast as desired.

2.5.3 Reading Environment

Reading rooms were all located in interior rooms with no windows. Doors were kept closed to eliminate extraneous light. Small task lamps were available in the room for those radiologists who wanted them. We did not measure the ambient light in the room with or without the task lamps on.

2.5.4 Reader Scoring

VIPER scoring was done per case, as in DMIST; there was no lesion localization. VIPER used a two-stage scoring system to allow radiologists to do their clinical task (recall the patient or not) and then be more quantitative. In the first stage, the readers were asked, "Would you recall this patient?" In the second stage of scoring, the readers were asked for a numerical score representing a likelihood or confidence the patient has cancer. The score is meant to rank patients; it is ordinal in nature and not meant to represent an actual probability or risk. The second stage yielded a 202-point ordinal score: 101 points for any no-recall decision and 101 points for any recall decision. The 101-point scale on either side of the threshold provides the radiologists ample space to be quantitative. By design, the two-stage scoring system yields a point on the empirical receiver operating characteristic (ROC) curve that exactly matches the sensitivity-specificity of the binary recall data.

We gave the readers instructions on how to navigate the electronic case report form and on how to provide the numeric score. These are available in the Supplementary Materials (17). The study administrator reviewed the instructions face-to-face with each reader before beginning data collection and then supervised the evaluation of four training cases.

2.6 Statistical Analyses

The primary analyses of VIPER were to compare reader-averaged empirical AUCs from FFDM and SFM for each reader study. We also compare reader-averaged sensitivities and specificities. Sensitivity and specificity use the recall decision, whereas AUC uses the 202-point ordinal score. The analyses were per case; there were no per-breast, lesion, or location analyses.

Each standard error (SE), p-value, and 95% confidence interval is estimated using U-statistic-based multiple-reader, multiple-case (MRMC) analysis methods, such that the analysis accounts for both reader and case variability (18,19). An MRMC analysis is expected in situations beyond exploratory studies (20) and in all ACRIN trials (21).

The p-values that we present are based on the standard inference test in which the null hypothesis assumes there is no performance difference between FFDM and SFM. The test statistic is the observed difference divided by the SE of that difference. The test statistic is assumed to follow a t-distribution for which we approximate the degrees of freedom (15). All the MRMC analyses were processed with version 4.0 of the iMRMC application developed at the FDA (22).

We sized the VIPER reader studies so that the SE of the differences in AUCs would be less than 0.03 (18,19). This would allow ample precision for testing an effect size of 0.11, the difference between the AUCs from FFDM and SFM observed in DMIST (1). The power for a significance level of 0.01 (0.05 split evenly among 5 reader studies) is 0.86. To do the sizing analysis, we used the variance components estimated from DMIST reader studies (23,24), which were obtained via ACRIN's data request mechanism. The MRMC analysis results of these studies, including the variance components used for sizing, can be found in the Supplementary Materials (17).

We created reader-averaged ROC curves by averaging the reader-specific non-parametric (trapezoidal) ROC curves along lines perpendicular to the chance line (25). This average is area-preserving; its AUC is equal to the reader-averaged nonparametric AUCs.

3. Results

Fig. 2 graphically compares the performance of FFDM to SFM for each VIPER reader study, including performance point-estimates and MRMC SE's. Refer to Table 3 for the MRMC analyses of reader-averaged performance differences: FFDM minus SFM. The FFDM and SFM reader-averaged ROC curves and operating points from VIPER are close. We are unable to reject any null hypothesis (at p=0.01) that there is no difference in the AUCs from FFDM and SFM. The SE for all AUC differences is less than 0.026, which meets the criteria that drove the study sizing.

The DMIST ROC curves and AUCs found in Fig. 2.D are reproduced from Fig. 1.C of the original DMIST results paper (1). They are based on pooling seven-point malignancy scores from all readers during DMIST screening, and then fitting a bivariate binormal model (26,27). The pooling mixes scores from different readers and can bias ROC curves and AUCs downward (28). We also caution that the DMIST estimates of SE do not account for reader variability due to the pooling across readers.

The DMIST sensitivities and specificities presented in Fig. 2.D are based on dichotomizing screening BIRADS scores (BIRADS 1,2,3 were negative; BIRADS 0,4,5 were positive). This dichotomization aligns with the VIPER dichotomization of "Recall" and "Do not recall". These sensitivities and specificities are reproduced from Table S2 of the DMIST Supplementary Materials (1).

On an absolute scale, the SFM AUCs in the VIPER screening studies (~0.73) are higher than the SFM AUC from DMIST (0.68), whereas the FFDM AUCs in the VIPER screening studies (~0.71) are lower than the FFDM AUC from DMIST (0.78). The uncertainties measured for AUCs in VIPER on these individual modalities are similar to those reported for DMIST, although the study designs, sample sizes, and variance estimation methods for the two studies are very different. The performance improvement with FFDM reported in DMIST for women with dense breasts is not replicated in any of the VIPER reader studies.

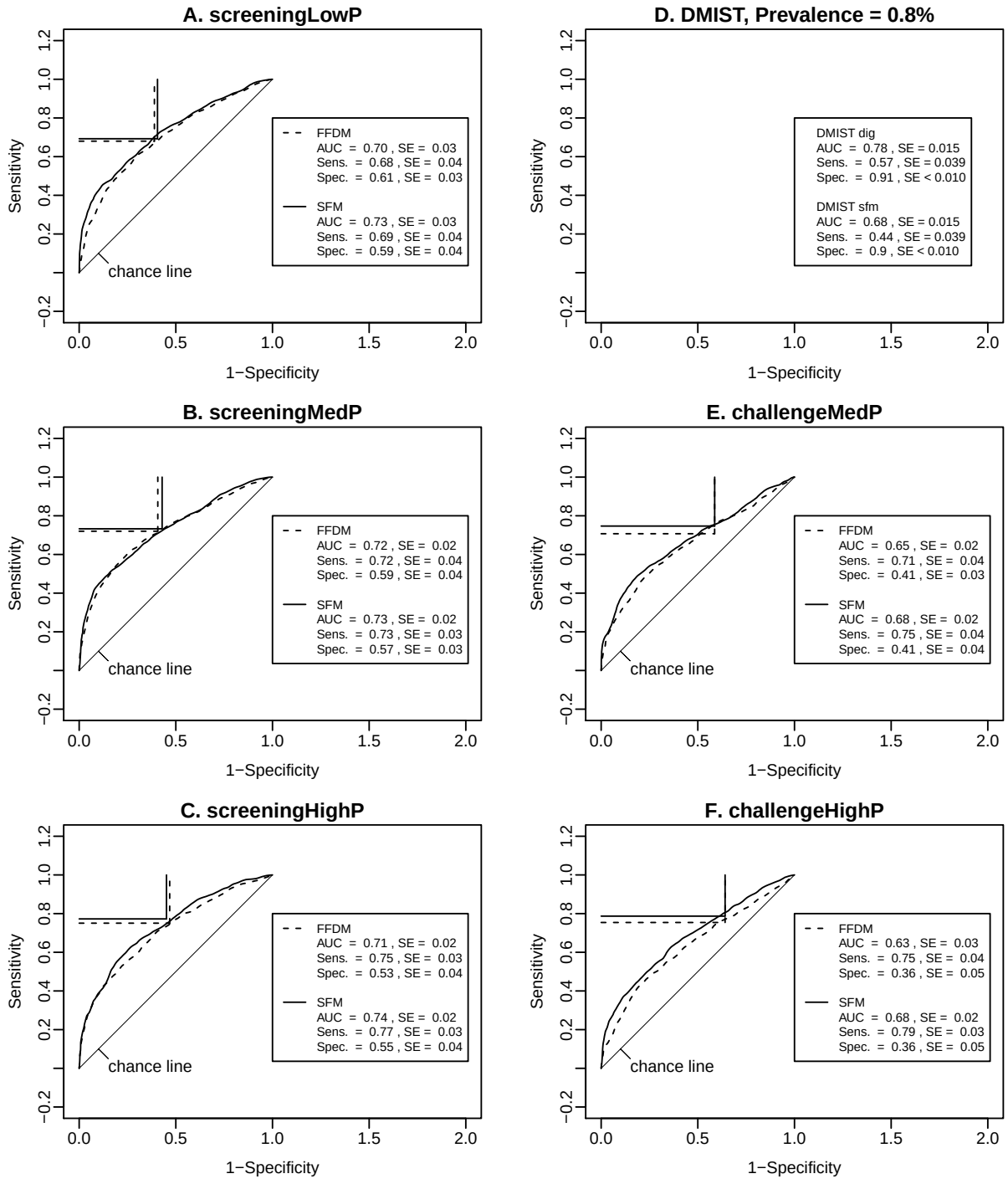

Figure 2: Figure 2: Plots of reader-averaged ROC curves and reader-averaged operating points (the vertical and horizontal crossings) for each of the VIPER reader studies. For each plot, we also provide the corresponding performance values and standard errors. In addition to the VIPER plots, we have added a reproduction of the related DMIST ROC curves in Plot D (reproduction from Fig. 1.C, "Women with Heterogeneously Dense" or "Extremely Dense Breasts" on Page 1778 of the DMIST NEMJ paper (1)) and the DMIST BIRADS sensitivity and specificity.

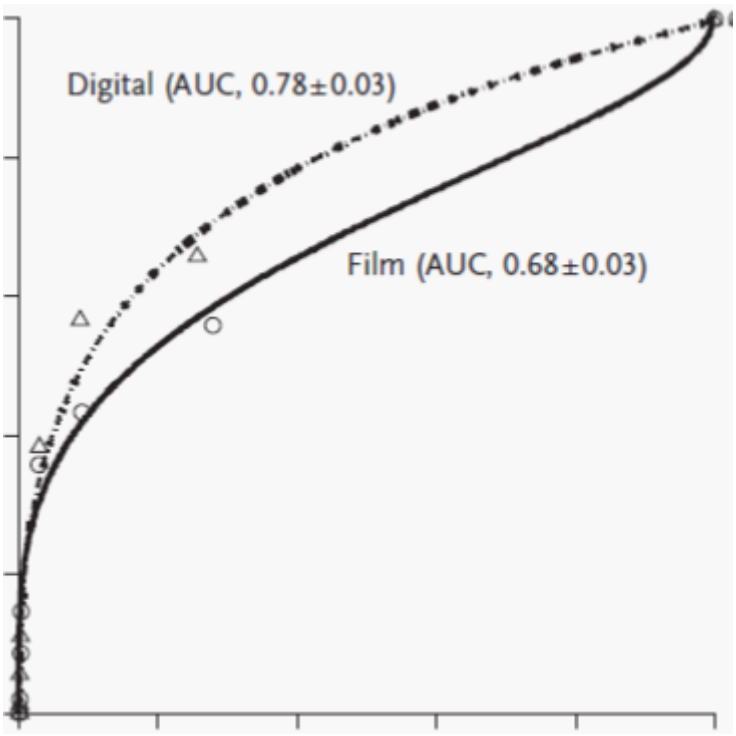

Figure 3: Figure 2D: This figure is sent separately from the rest of Figure 2 as it comes from a separate source. Please crop out everything but the axes and the content inside the axes and align it with the axes provided for Figure 2D in the main figure.

Fig. 3 compares performance as a function of prevalence for each combination of modality (FFDM and SFM) and study population (screening and challenge). In each plot, prevalence does not appear to be impacting the ROC curves. The DMIST operating point for women with dense breasts is also provided on these plots for reference.

The reader-specific operating points in Fig. 3 are widely dispersed. Specificity ranges from below 0.1 to 0.8 and sensitivity ranges from 0.4 to 1.0. This data demonstrates reader variability in skill and threshold on the clinical decision to "recall" and "do not recall" in the VIPER lab-based study conditions. When we average the operating points over the readers, they trend up and to the right as prevalence is increased in each sub-plot of Fig. 3. The trend is moderate and needs to be explored in a statistical model aggregating the results across all the VIPER reader studies. The trend is consistent with the expected behavior of a decision-maker that is maximizing a risk-benefit relationship between the true and false positives, and the true and false negatives (18). When we compare the reader-averaged VIPER operating points to those from DMIST (where prevalence was much lower, 0.8%), it is clear that the VIPER operating points are above and to the right; readers are more aggressive calling cases positive in VIPER.

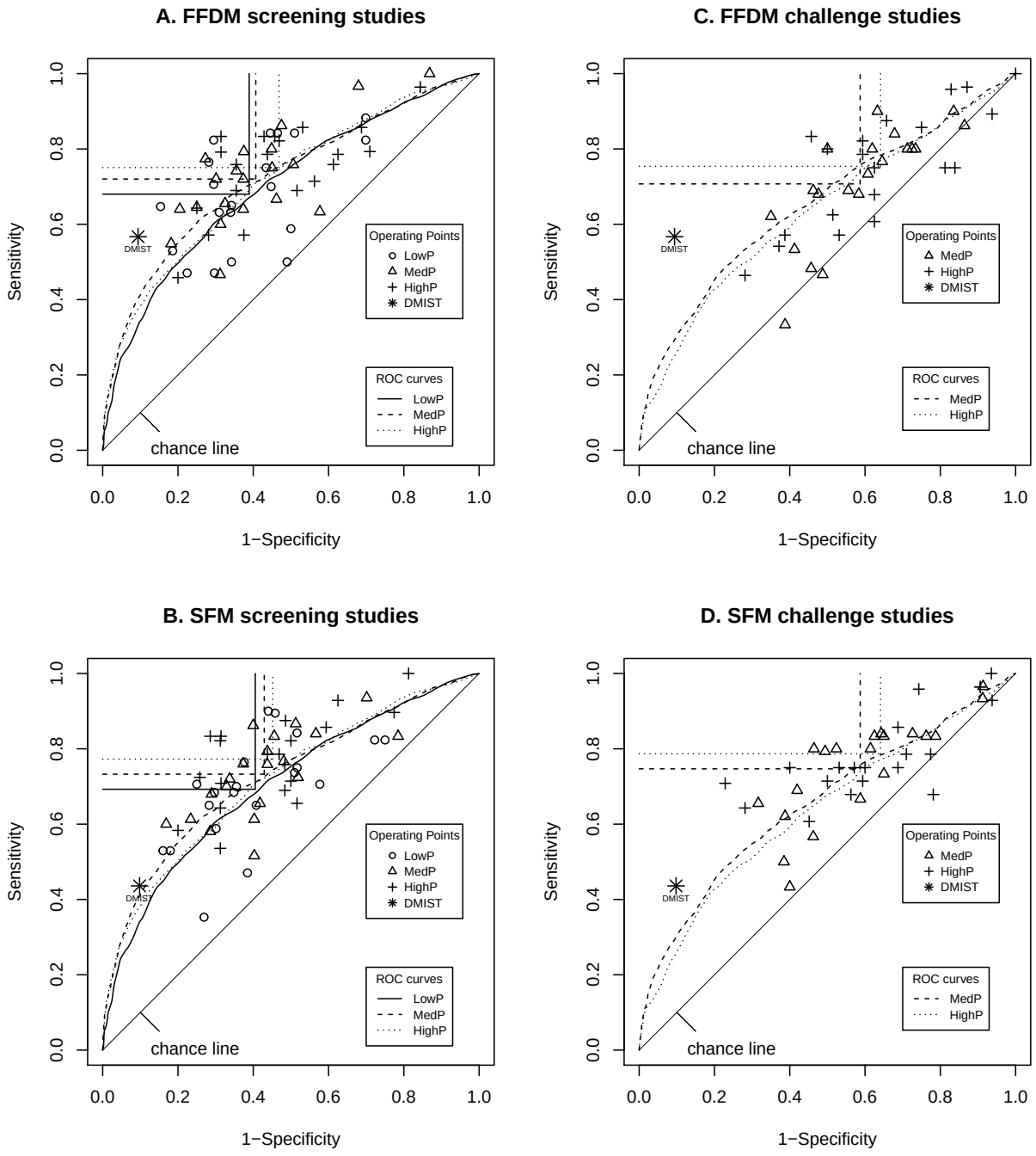

Figure 4: Figure 3: Plots of reader-averaged ROC curves, reader-averaged (1-Spec., Sens.) operating points (the vertical and horizontal crossings), and reader-specific operating points (denoted by the symbols). Study populations are restricted to women with dense breasts (Heterogeneously Dense and Extremely Dense). Reader-averaged ROC curves of different prevalences are very close. Reader-averaged operating points move up and to the right as prevalence increases.

Table 3 title: Table of MRMC performance differences for AUC, sensitivity, and specificity.

Table 3 footnote: We are unable to reject any null hypothesis that there is no difference in the AUCs from FFDM and SFM with a significance level of 0.01 (0.05 split evenly between 5 reader studies). Confidence intervals are not reduced to account for multiplicity. Individual modality performance results can be found in Fig. 2.

Note: Table 3 is actually 3 concatenated tables: one for AUC, Sensitivity, and Specificity."

Reader Study (AUC)	Prevalence (%)	Number of Observations	Difference	Standard Error	95% Confidence Interval
screeningLowP	10.6	6911	-0.029	0.024	(-0.078, 0.021)
screeningMedP	26.6	4325	-0.005	0.024	(-0.054, 0.043)
screeningHighP	45.6	2390	-0.025	0.025	(-0.075, 0.024)
challengeMedP	26.1	4377	-0.024	0.018	(-0.06, 0.013)
challengeHighP	45.2	2379	-0.047	0.023	(-0.093, -0.001)
DMIST	0.8	39794	0.110	0.035	(0.04, 0.18)

Reader Study (Sensitivity)	Prevalence (%)	Number of Observations	Difference	Standard Error	95% Confidence In
screeningLowP	10.6	730	-0.013	0.033	(-0.08, 0.055)
screeningMedP	26.6	1148	-0.012	0.033	(-0.081, 0.056)
screeningHighP	45.6	1090	-0.022	0.025	(-0.073, 0.029)
challengeMedP	26.1	1140	-0.040	0.025	(-0.09, 0.011)
challengeHighP	45.2	1077	-0.033	0.029	(-0.093, 0.027)
DMIST	0.8	330	0.131	0.047	(0.034, 0.222)

Reader Study (Specificity)	Prevalence (%)	Number of Observations	Difference	Standard Error	95% Confidence Int
screeningLowP	10.6	6181	-0.016	0.015	(-0.013, 0.045)
screeningMedP	26.6	3177	-0.022	0.022	(-0.023, 0.067)
screeningHighP	45.6	1300	0.017	0.031	(-0.081, 0.047)
challengeMedP	26.1	3237	0.000	0.017	(-0.035, 0.035)
challengeHighP	45.2	1302	0.000	0.030	(-0.061, 0.06)
DMIST	0.8	39464	0.004	0.003	(-0.001, 0.01)

4. Discussion

VIPER found consistent AUC comparisons of FFDM and SFM across different case distributions and prevalence. We also found that AUC appears robust to changes in prevalence, whereas sensitivity and specificity appear to depend on prevalence. This is consistent with the literature mentioned in the Introduction (3–10).

However, VIPER did not find a statistically significant difference in performance between FFDM and SFM for women with dense breasts. This conflicts with the statistically significant difference found in the DMIST prospective clinical trial. We attribute this conflicting result to the different statistical analyses and the different image processing. The VIPER studies also found very different results for sensitivity and specificity compared to DMIST. Clearly the behavior of the radiologists is different in these lab-based studies compared to the (very low prevalence) prospective trial. Specifically, the radiologists' sensitivity and specificity in VIPER reflect an increased level of aggressiveness in recalling patients compared to DMIST (See Fig. 3). Compared to what has been documented before, this is a more dramatic difference on a clinically relevant task across five studies and a very large prospective trial.

To our knowledge no clinical studies other than Pisano et al. have published AUC results comparing FFDM to SFM on women with dense breasts. Kerlikowske et al. 2011 (29) found sensitivity to be similar for women with heterogeneously dense breasts (FFDM: 82%, SFM: 79%, 1069 cancers) and borderline significantly higher for FFDM for women with extremely dense breasts (FFDM: 84%, SFM: 68%, 163 cancers). At the same time, specificity was significantly lower for FFDM (FFDM: 87%, SFM: 90%, 240,756 non-cancers). Similarly, Del Turco et al. 2007 (30) found a higher cancer detection rate for FFDM compared to SFM (FFDM: 1.05%, SFM: 0.53%) that was accompanied by a higher recall rate (FFDM: 4.85%, SFM: 2.69%) for women with a breast density larger than 75%. It is not known if these differences are related to differences in AUCs; the differences could be due to a difference in device performance or a change in decision threshold. In VIPER, we were able to elicit ROC scores on a 202-point scale that sampled all of ROC space without gaps, eliminating ambiguity that comes from comparing operating points. Please see the individual reader ROC curves in the Supplementary Materials (17).

The VIPER and DMIST study designs were quite different. Most of the differences were the effects under investigation; the differences were by design. VIPER was retrospective, not prospective. There was no patient information used during the image interpretation (including prior images), and the image interpretation did not affect patient management. Without patient management, radiologists may be less vigilant. Regarding cases, the patient distribution was out of balance in terms of subpopulations (we only included women with dense breasts), prevalence (we enriched with cancer cases), and case mix (we investigated challenging populations in addition to a screening population). These differences were clear at the beginning of the study to the designers as well as the participants and, except for the focus on the subpopulation of women with dense breasts, the differences were not expected to favor one modality over the other.

VIPER utilized a split-plot study design, which can make efficient use of cases, each reader's time, and the total number of observations of a study (15,16). Each case is read by multiple readers, reducing the noise from a single observation. Each case is not read by all readers, avoiding the diminishing returns from adding too many reads. Ultimately, the split-plot study design allowed us to efficiently utilize the limited cancer cases available to us and to conduct five split-plot studies for the cost of two fully-crossed studies without sacrificing statistical precision, even in the low-prevalence reader study.

At the time of the design of the DMIST study, there were no validated and readily available MRMC analysis methods for the DMIST data, methods that accounted for reader and case variability in study designs that were not fully crossed. Consequently, the DMIST study pooled the results over readers; pooling effectively ignores reader variability and can bias ROC curves and AUC downward. There are now methods that can perform MRMC analyses of alternate study designs (15,19,31,32), and some of those methods were used for this work. Consequently, caution should be taken when comparing the pooled analyses from DMIST to the MRMC analyses here.

The most critical limitation of VIPER is that the image processing and display were not the same as was done in DMIST. The DMIST study had staff, including manufacturers' engineers, dedicated to optimizing display, providing calibration and quality control. The original image processing algorithms applied during

DMIST were not available for this study. VIPER only had access to the images in the raw format, and they were all processed in the same way, regardless of image characteristics or vendor. Furthermore, the monitors were not calibrated.

In closing, we found that AUC differences in lab-based reader studies were robust to changes in the study population (prevalence and the distribution of non-cancer cases), and the split-plot study design was effective at reducing the number of observations needed in total and per reader. In comparing to DMIST, we feel that it was important to have multiple readers evaluate the rare cancer cases, reducing the impact that one reader can have on the study results. As such, we caution investigators to design their studies so that each reader evaluates an adequate number of positive and negative cases. It should be possible to calculate each reader's AUC performance, and statistical analyses should account for reader and case variability. Following others, we think a useful rule of thumb is to have at least 20 positive and 20 negative cases per reader (15). While lab-based reader studies can be designed with these minimums in mind, prospective studies might require special methods, like a feedback loop to send rare diseased cases to other readers (only possible in very large studies). Regardless, we also recommend that the workload be balanced across readers for lab-based and prospective studies; each reader should evaluate a similar number of cases. When readers read different caseloads, the results depend on how the interpretations of different readers are combined (the weights) and there is more reader variability in the results (31).

The VIPER data, analysis scripts, and Supplementary Materials are available online (17).

Acknowledgements

We thank Norberto Pantoja-Galicia and all our colleagues in the FDA CDRH Office of Science and Engineering Laboratories and the Office of Surveillance and Biostatistics for their feedback on the study design, analyses, and results. We would also like to acknowledge Qi Gong and his work cleaning the data for processing and supporting the data analysis.

Contributors

All authors contributed to the conception or design of the work, the acquisition, analysis, or interpretation of the data. All authors were involved in drafting and commenting on the paper and have approved the final version.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Disclosures

As employees of the Medical University of South Carolina, Dr. Etta Pisano and Ms. Elodia Cole were contracted by the FDA to execute the study (access and prepare the images from ACRIN, recruit and pay readers, and host the data collection). Dr. Pisano is the only author with potential conflicts of interests. Her potential conflicts of interest are

- Freenome Holdings, Inc. July 2018-June 2019
- Real Imaging April 2018-May 2020
- Therapixel July 2018-June 2019
- Phillips Corporation - Contract w/my employer for research (Ended 6/30/17)
- Fuji Film - Contract w/my employer for research (Ended 6/30/17)
- Alan Penn Associates - SBIR subcontract (Ended 8/31/17)

- DEEPHEALTH, INC. - Advisory Board member (Ended 1/27/18)

All other authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval

Informed consent was obtained from all radiologist participants. The Medical University of South Carolina institutional review board approved the study (IRB 13890).

Biographies

** Brandon D. Gallas ** provides mathematical, statistical, and modeling expertise to the evaluation of medical imaging devices at the FDA. His main areas of contribution are in the design and statistical analysis of reader studies (https://github.com/DIDSR/iMRMC/releases, https://cran.r-project.org/web/packages/iMRMC/index.html), image quality, computer-aided diagnosis, and imaging physics. Before working at the FDA, he was in Dr. Harrison Barrett's radiology research group at the University of Arizona earning his PhD degree in applied mathematics from the Graduate Interdisciplinary Program.

** Weijie Chen ** received his PhD in medical physics from the University of Chicago, Chicago, Illinois, in 2007. Since then, he has been a scientist at the Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland. His research interests include statistical assessment methodologies for diagnostic devices in general, and evaluation of medical imaging, artificial intelligence algorithms, and computer-aided diagnosis systems in particular.

** Elodia Cole ** is a biomedical imaging researcher who has specialized in breast imaging technology clinical effectiveness research for 18 years. She graduated with a Master of Science degree in Biomedical Engineering from UNC-CH in 2000. Her research has included the application of various image processing, image display, and image analysis tools and their impact on radiologist performance in the detection of breast cancer.

** Dr. Robert Ochs ** is the Deputy Director for Radiological Health within the Food and Drug Administration, Center for Devices and Radiological Health, Office of In Vitro Diagnostics and Radiological Health. His office is responsible for the pre-market review and regulation of radiological medical devices, regulation of radiation emitting electronic products, and implementation of the Mammography Quality Standards Act. He received his Ph.D. in Biomedical Physics from the University of California, Los Angeles.

** Nicholas Petrick ** is an SPIE Fellow and is Deputy Director for the Division of Imaging, Diagnostics and Software Reliability at the Center for Devices and Radiological Health, U.S. Food and Drug Administration. This Division conducts regulatory research in medical imaging physics and image analysis. Dr. Petrick received his Ph.D. from the University of Michigan and his current research interests include quantitative imaging, medical machine learning and assessment methods for medical imaging and machine learning devices.

** Dr. Etta Pisano ** is Professor in Residence of Radiology at Harvard Medical School and Chief Research Officer at the American College of Radiology. She is also currently serving as the Study Chair for The Tomosynthesis Mammographic Imaging Screening Trial, a National Cancer Institute-funded clinical trial under the auspices of the ECOG-ACRIN research base, which will enroll over 165,000 women in 150 sites in the US and Canada.

** Berkman Sahiner ** has a PhD in electrical engineering from the University of Michigan. He was an associate professor at the Radiology Department, the University of Michigan, until 2009. Since then, he has been a senior biomedical research scientist at the U.S. Food and Drug Administration. His research interests include computer-aided diagnosis, machine learning, image analysis, breast imaging, image perception, and performance assessment methodologies.

** Frank Samuelson ** works as a physicist at the US Food and Drug Administration. His research includes methods of evaluating computational intelligence algorithms found in diagnostic medical devices, such as statistical methods and study designs. He is an expert in evaluating signal detection involving human observers, and he reviews studies for devices and algorithms for the FDA.

** Kyle J. Myers ** received a Ph.D. in Optical Sciences from the University of Arizona in 1985. She is the Director of the Division of Imaging, Diagnostics, and Software Reliability, FDA/CDRH. She and Dr. Harrison H. Barrett coauthored Foundations of Image Science (2004), winner of the First Biennial J.W. Goodman Book Writing Award from OSA and SPIE. Dr. Myers is a Fellow of AIMBE, OSA, SPIE, and a member of the National Academy of Engineering.

Table Captions

Table 1 title: Table of average per-reader cancer prevalence.

Table 1 footnote: Cancer cases are those that were pathologically verified within 455 days after the initial study mammogram. The average number of cancers and non-cancers are presented here as the case subgroups were not all equal. An "observation" is one radiologist's score for one case. VIPER reader studies only included women with heterogeneously dense or extremely dense breasts. For reference, pooling over all DMIST radiologists, DMIST found 165 cancers in 19,897 women with dense breasts (prevalence: 0.8%).

Table 2 title: Table of per-reader distributions of BIRADS patient management scores for women without cancer.

Table 2 footnote: Average counts are per reader and based on DMIST SFM and FFDM evaluations. Average counts are presented as the case subgroups were not all equal. For reference, DMIST SFM screening yielded 39,013 BIRADS 1 & 2 women compared to 3648 BIRADS 0 women (ratio: 10.69).

Table 3 title: Table of MRMC performance differences for AUC, sensitivity, and specificity.

Table 3 footnote: We are unable to reject any null hypothesis that there is no difference in the AUCs from FFDM and SFM with a significance level of 0.01 (0.05 split evenly between 5 reader studies). Confidence intervals are not reduced to account for multiplicity. Individual modality performance results can be found in Fig. 2.

Figure Captions

Figure 1: Participant flow diagram. Screening and Challenge sub-studies of different prevalences are created from the final subgroups based on DMIST screening BIRADS scores by FFDM and SFM. The last row differs from the one above it due to availability from ACRIN. (Figure width = page.)

Figure 2: Plots of reader-averaged ROC curves and reader-averaged operating points (the vertical and horizontal crossings) for each of the VIPER reader studies. For each plot, we also provide the corresponding performance values and standard errors. In addition to the VIPER plots, we have added a reproduction of the related DMIST ROC curves in Plot D (reproduction from Fig. 1.C, "Women with Heterogeneously Dense" or "Extremely Dense Breasts" on Page 1778 of the DMIST NEMJ paper (1)) and the DMIST BIRADS sensitivity and specificity. (Figure width = page.)

Figure 3: Plots of reader-averaged ROC curves, reader-averaged (1-Spec., Sens.) operating points (the vertical and horizontal crossings), and reader-specific operating points (denoted by the symbols). Study populations are restricted to women with dense breasts (Heterogeneously Dense and Extremely Dense). Reader-averaged ROC curves of different prevalences are very close. Reader-averaged operating points move up and to the right as prevalence increases. (Figure width = page.)

References

1. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med. 2005;353(17):1773–1783.

2. Pisano ED, Gatsonis CA, Yaffe MJ, et al. American college of radiology imaging network digital mammographic imaging screening trial: Objectives and methodology. Radiology. 2005;236(2):404–412.

3. Ulehla ZJ. Optimality of perceptual decision criteria. J Exp Psychol. 1966;71(4):564–569.

4. Wolfe JM, Horowitz TS, Kenner NM. Cognitive psychology: Rare items often missed in visual searches. Nature. Visual Attention Laboratory, Brigham; Women's Hospital, Boston, Massachusetts 02139, USA. wolfe@search.bwh.harvard.edu; 2005;435(7041):439–440http://dx.doi.org/10.1038/435439a.

5. Egglin TKP, Feinstein AR. Context bias: A problem in diagnostic radiology. JAMA. 1996;276:1752–1755.

6. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. Radiology. 2003;228(1):10–14.

7. Gur D, Rockette HE, Warfel T, Lacomis JM, Fuhrman CR. From the laboratory to the clinic: The "prevalence effect". Acad Radiol. Department of Imaging Research, Suite 4200, 300 Halket Street, University of Pittsburgh, Pittsburgh, PA 15213-3180, USA. 2003;10(11):1324–1326.

8. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in a laboratory environment: Changing the confidence ratings. Acad Radiol. Department of Radiology, Imaging Research, Suite 4200 Magee-Womens Hospital, 300 Halket Street, School of Medicine, Pittsburgh, PA 15213-3180, USA. gurd@upmc.edu; 2007;14(1):49–53http://dx.doi.org/10.1016/j.acra.2006.10.003.

9. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology. Department of Radiology, University of Pittsburgh School of Medicine, 3362 Fifth Ave, Pittsburgh, Pa 15213-31803, USA. gurd@upmc.edu; 2008;249(1):47–53http://dx.doi.org/10.1148/radiol.2491072025.

10. Evans KK, Tambouret RH, Evered A, Wilbur DC, Wolfe JM. Prevalence of abnormalities influences cytologists' error rates in screening for cervical cancer. Arch Pathol Lab Med. 2011;135(12):1557–1560.

11. Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: Why some cancers are missed in breast cancer screening. PloS One. 2013;8(5):e64366.

12. Gallas BD, Pisano E, Cole E, Myers K. Impact of different study populations on reader behavior and performance metrics: Initial results. Kupinski MA, Nishikawa RM, editors. Proc SPIE. 2017;10136:0A.

13. D'Orsi CJ, Mendelson EB, Ikeda DM, et al. Breast imaging reporting and data system (bi-rads). 4th ed. Reston, VA: American College of Radiology; 2003.

14. Sickles EA, D'Orsi CJ, al. LWB et. ACR bi-rads mammography. ACR bi-rads atlas, breast imaging reporting and data system. 5th ed. Reston, VA: American College of Radiology; 2013.

15. Obuchowski N, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: A comparison of statistical methods. Acad Radiol. 2012;19(12):1508–1517.

16. Chen W, Gong Q, Gallas BD. Paired split-plot designs of multireader multicase studies. Journal of Medical Imaging. 2018;5:031410https://doi.org/10.1117/1.JMI.5.3.031410.

17. Gallas BD, Chen W, Cole E, et al. Supplementary materials: Impact of prevalence and case distribution in lab-based diagnostic imaging studies. 2018.https://github.com/DIDSR/iMRMC/wiki/Gallas2018_VIPER. Accessed December 30, 2018.

18. Gallas BD, Bandos A, Samuelson F, Wagner RF. A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators. Commun Stat A-Theory. 2009;38(15):2586–2603.

19. Gallas BD, Brown DG. Reader studies for validation of CAD systems. Neural Networks Special Confer-

ence Issue. 2008;21(2):387–397http://www.sciencedirect.com/science/article/pii/S0893608007002456.

20. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: Is there a better way? Radiology. 2010;257(1):14–17http://dx.doi.org/10.1148/radiol.10100252.

21. Hillman BJ. ACRIN–lessons learned in conducting multi-center trials of imaging and cancer. Cancer Imaging. Department of Radiology, University of Virginia, Charlottesville, Virginia, USA. bjh8a@virginia.edu; 2005;5 Spec No A:S97–101http://dx.doi.org/10.1102/1470-7330.2005.0026.

22. Gallas BD. IMRMC v4.0: Application for analyzing and sizing MRMC reader studies. Silver Spring, MD: Division of Imaging, Diagnostics,; Software Reliability, OSEL/CDRH/FDA; 2017.https://github.com/DIDSR/iMRMC/releas https://cran.r-project.org/web/packages/iMRMC/index.html. Accessed January 15, 2018.

23. Nishikawa RM, Acharyya S, Gatsonis C, et al. Comparison of soft-copy and hard-copy reading for full-field digital mammography. Raidology. 2009;251(1):41–51.

24. Hendrick RE, Cole EB, Pisano ED, et al. Accuracy of soft-copy digital mammography versus that of screen-film mammography according to digital manufacturer: ACRIN dmist retrospective multireader study. Radiology. hwestern Univ, Galter Pavilion, 13th Floor, 251 E Huron St, Chicago, IL 60611, USA. edward.hendrick@gmail.com; 2008;247(1):38–48http://dx.doi.org/10.1148/radiol.2471070418.

25. Chen W, Samuelson FW. The average receiver operating characteristic curve in multi-reader multi-case imaging studies. Br J Radiol. Division of Imaging; Applied Mathematics Office of Science; Engineering Laboratories Center for Devices; Radiological Health Food; Drug Administration 10903 New Hampshire Avenue Silver Spring, Maryland, 20993 United States. 2014;87:20140016http://dx.doi.org/10.1259/bjr.20140016.

26. Metz CE, Wang P -L., Kronman KB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconick F, editor. Information processing in medical imaging VIII. Springer, Netherlands; 1984. pp. 432–445.

27. Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. Second. Hoboken, New Jersey: Wiley & Sons; 2011.

28. Pepe MS. The statistical evaluation of medical tests for classification and prediction. UK: Oxford University Press; 2003.

29. Kerlikowske K, Hubbard RA, Miglioretti DL, et al. Comparative effectiveness of digital versus film-screen mammography in community practice in the united states: A cohort study. Ann Intern Med. University of California, San Francisco, USA. Karla.Kerlikowske@ucsf.edu; 2011;155(8):493–502.

30. Del Turco MR, Mantellini P, Ciatto S, et al. Full-field digital versus screen-film mammography: Comparative accuracy in concurrent screening cohorts. Am J Roentgenol. 2007;189(4):860–866.

31. Gallas BD, Pennello GA, Myers KJ. Multireader multicase variance analysis for binary data. J Opt Soc Am A, Special Issue on Image Quality. 2007;24(12):B70–B80.

32. Hillis SL. A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data. Stat Med. 2014;33(2):330–360http://dx.doi.org/10.1002/sim.5926.