FDA Executive Summary

Meeting of the Radiological Devices Advisory Panel

On September 24, 2010, the committee will discuss, make recommendations, and vote on a premarket approval application for the Selenia Dimensions 3D* digital breast tomosynthesis (DBT) system (P080003), sponsored by Hologic, Inc. The Selenia Dimensions 3D system is intended for use in the same clinical applications as 2D full field digital mammography (FFDM). The screening examination is intended to consist of a 2D FFDM image set plus a 3D DBT image set.

Federal Notice
http://edocket.access.gpo.gov/2010/2010-18416.htm

Panel Date and Location
September 24, 2010 from 8 AM to 6 PM
Hilton Washington DC North/Gaithersburg
620 Perry Pkwy., Gaithersburg, MD 20877

Contact Person
Shanika Craig
Food and Drug Administration
Center for Devices and Radiological Health
10903 New Hampshire Ave., Bldg 66, Room 1613
Silver Spring, MD 20993-0002
301-796-6639

Executive Summary Authors
Kish Chakrabarti, PhD
Robert Ochs, PhD
Gene Pennello, PhD
Frank Samuelson, PhD

*The official device name on the P080003 submission was for the Selenia C Full Field Digital Mammography (FFDM) system. The planned trade name for the device is the Selenia Dimensions (2D and 3D), where 2D refers to conventional 2D FFDM and 3D refers to the new 3D Digital Breast Tomosynthesis (DBT). For clarity, the executive summary refers to the 2D FFDM plus 3D DBT system as the "Selenia Dimensions 3D" or just "2D plus 3D".

Table of Contents

Section **Page Number**

Introduction 3

I. Rationale for Presentation to the Panel 5
A. Study design 5
B. Screening claim 5
C. Diagnostic claim 6

II. Background Information 7
A. Indications for Use 7
B. Device Description 7
C. Non-clinical Testing 7
i. Device Characteristics 7
ii. Performance Standards 7
iii. Technical Data 8
D. Radiation Dose 8

III. Reader Studies 9
A. Reader Study 1 (Pivotal Study) 9
i. Study Design 9
iii. Results 11
B. Reader Study 2 20
i. Study Design 20
ii. Results 21
C. Pittsburgh Reader Study 26
i. Abstract 26
ii. Analysis 27
D. Discussion 29

IV. Panel Questions 33
A. Questions for Panel Discussion 33
B. Questions for Ballot Vote 34

Introduction

The subject of this Executive Summary is the Selenia Dimensions 3D system premarket approval (PMA, P080003) application, sponsored by Hologic, Inc. Your time and effort in review of this application is greatly appreciated.

The Selenia Dimensions system is a first-of-a-kind device capable of acquiring both conventional 2D full field digital mammography (FFDM) and 3D digital breast tomosynthesis (DBT) images. The Selenia Dimensions system is intended for use in the same clinical applications as 2D FFDM.

This executive summary discusses three multi-reader multi-case (MRMC) studies that compared 2D FFDM alone to 2D FFDM plus 3D DBT. Hologic performed two studies (Reader Study 1 and Reader Study 2) and a third study was performed by the University of Pittsburgh.

Reader Study 1 was the pivotal study of the PMA. Hologic submitted Reader Study 2 and the Pittsburgh study in response to deficiencies from FDA. In particular, the results of Reader Study 2 were provided to support a lower dose tomosynthesis protocol (i.e., 2D plus 3D MLO) as well as address concerns with the reader scoring methodology (i.e., identification of the correct location of a malignant lesion was not required for crediting readers with a true positive result) and impact of 2D plus 3D on the recall rate of cancer patients. Reader Study 2 used new readers and a new random selection of non-cancer cases. Reader Study 2 reused the cancer cases from Reader Study 1 with the addition of three more cancers.

Hologic also submitted a research study performed by the University of Pittsburgh. The research study is relevant to the discussion, but the study was not designed or intended for the FDA review in this PMA submission. The study used different cancer cases than Reader Study 1 and Reader Study 2.

Table 1: Summary of the 2D/3D notation used in the executive study. "2D alone" refers to a set of conventional 2-view FFDM images; a cranio-caudal (CC) and mediolateral oblique (MLO) view for each breast. "2D plus 3D MLO" refers to the conventional 2D FFDM plus 1 MLO tomosynthesis view. "2D plus 3D" refers to conventional 2D FFDM plus 2 DBT views (MLO and CC).

Mode	Images	Reader Study 1	Reader Study 2	Pittsburgh Study
2D alone	2-view FFDM (MLO and CC)	Yes	Yes	Yes
2D plus 3D MLO	2-view FFDM (MLO and CC), and 1-view DBT (MLO)	-	Yes	-
2D plus 3D	2-view FFDM (MLO and CC), and 2-view DBT (MLO and CC)	Yes	Yes	Yes

This executive summary includes four sections:

1. Summary of the issues that FDA would like to bring to the attention of the panel;
2. Summary of the Indications for Use, device operation, and non-clinical testing;
3. Review of the clinical reader studies; and
4. Questions for panel discussion and voting.

I. Rationale for Presentation to the Panel

FDA seeks advice from the Radiological Devices Panel on the potential risk versus benefit of using 2D FFDM plus 3D DBT for screening and diagnosis of breast cancer. The panel is asked to consider the results from the pivotal study (Reader Study 1), the two supplemental reader studies, the study design, and the degree of generalizability to the intended populations. FDA would also like the panel to consider whether the potential benefit of 3D tomosynthesis outweighs the risk from the additional radiation exposure.

A. Study design

Because a prospective study is prohibitive due to low prevalence of cancer, retrospective studies were conducted that were enriched with cancer cases. The studies conducted have the following limitations, which might affect the generalizability of the results to the intended population:

- Radiologists are reading images knowing that their readings do not affect patient management;
- Radiologists are scoring the image without the use of prior images or clinical history;
- Enrichment with recalls and cancer cases that were identified based on standard imaging, including FFDM, may bias the study against 3D relative to 2D FFDM;
- Radiologists know or will learn that the case set is highly enriched with cancer cases;
- Making multiple readings successively in different modes (sequential reading) potentially confounds modality effects with the effects of having additional reading time; and
- Ground truth is not established definitely for normal cases (BIRADS=0) because they were not followed up to confirm that they did not have cancer.

In addition, prior to the second Hologic study (Reader Study 2), the training was enhanced based on the types of errors made during Reader Study 1; this also raises concern as the same cancer cases were used in both studies.

B. Screening claim

In both Reader Study 1 and Reader Study 2, use of 2D FFDM plus 3D DBT significantly increased the area under the ROC curve (AUC) and significantly decreased the recall rate of non-cancer patients.

In Reader Study 1, the model-based AUC estimate increased from 0.821 for 2D to 0.894 for 2D plus 3D for a difference of 0.072 (p-value 0.0001). In Reader Study 2, the model-based AUC estimate increased from 0.828 for 2D to 0.895 for 2D plus 3D for a difference of 0.068 (p-value < 0.0001).

The primary objective for recall rate was to show a significant reduction in the recall rate among the screening cases. In Reader Study 1, the average recall rate for screening cases went from 51.5% with 2D to 12.9% with 2D plus 3D, and in Reader Study 2, the average recall rate for screening cases went from 44.2% with 2D to 24.0% with 2D plus 3D.

In Reader Study 1, the average recall rate of ***cancer*** patients also decreased (from 87.2% to 80.4%); the decrease in the recall rate of cancer patients was significant for 4 of 12 readers. In Reader Study 2, the recall rate of cancer patients remained the same, at 88%, on average across readers.

C. Diagnostic claim

In addition to the primary endpoints, the sponsor provided a "Forced BIRADS" task based comparison of using 2D alone with using 2D plus 3D tomosynthesis in order to support the intended use of the device for "screening and diagnosis".

FDA would like the panel to consider whether there is sufficient data to support 3D tomosynthesis for use as another exam option in the diagnostic setting (i.e., the "…and diagnosis" part of the intended use).

II. Background Information

A. Indications for Use

The Hologic Selenia Dimensions System generates digital mammographic images that can be used for screening and diagnosis of breast cancer. The Selenia Dimensions (2D or 3D) system is intended for use in the same clinical applications as 2D mammography systems for screening mammograms. Specifically, the Selenia Dimensions system can be used to acquire 2D digital mammograms and 3D mammograms. The screening examination will consist of a 2D image set or a 2D and 3D image set. The Selenia Dimensions system may also be used for additional diagnostic workup of the breast.

B. Device Description

The Selenia Dimensions 3D DBT is a hardware and software upgrade to the Selenia Dimensions 2D FFDM system, which is FDA approved for conventional mammography imaging (P010025/S013, approved December 22, 2008).

To acquire the 3D DBT images, the x-ray tube head is moved in a 15^{o} arc over the stationary breast acquiring 11 to 21 (typically 15) low dose x-ray projection images. The projection images are reconstructed to produce cross-sectional "slices" through the breast. The nominal thickness of the slices can vary from 0.5 to 10 mm, with 1 mm being the "normal" thickness.

The same detector and x-ray tube is used to acquire both the 2D and 3D images. Images can be acquired in any orientation of the gantry, including the standard CC and MLO mammography views. The 2D and 3D images can be acquired during a single breast compression, or they can be acquired separately.

C. Non-clinical Testing

i. Device Characteristics

Hologic provided a sufficient description of the entire system, including the fundamental components, and principles of operation. The sponsor also compared the PMA device to the prototype used in the clinical trial.

ii. Performance Standards

The sponsor provided certification that the device conforms to the regulations issued pursuant to the Radiation Control for Health and Safety Act of 1968. The sponsor also stated compliance with 21 CFR Part 820 (Quality Systems Regulations), ISO 13485 (Medical Devices – Quality management systems – Requirements for regulatory purposes), ACR/NEMA Digital Imaging Communications in Medicine standard (DICOM), and several voluntary IEC standards for the safety and performance of medical electrical equipment.

iii. Technical Data

The sponsor provided experimentally measured data, including: sensitometry, spatial resolution as described by the modulation transfer function (MTF), and SNR transfer, as described by detective quantum efficiency (DQE). The sponsor provided the results of imaging experiments using multiple phantoms, including the FDA approved ACR accreditation phantom and the CD-MAM phantom. The sponsor used phantoms of thickness 2-6 cm to estimate the average glandular dose for exposures made with the Automatic Exposure Control (AEC) system with varying breast phantom compositions (i.e., 70/30, 50/50, and 30/70) of glandular and adipose tissue, respectively. The sponsor used the methodology described in "Contrast Detail Phantom Scoring Methodology", Thomas, Chakrabarti, Kaczmarek and Romanyukha, Med. Phys. Vol. 32, No. 3, p.807-814, 2005. The sponsor provided test information on both the prototype used for the clinical study and the Selenia Dimensions 3D system.

The Selenia Dimensions 3D PMA application does not include the display system or review workstation. There are no special hardware requirements for the workstation to support the review of tomosynthesis images beyond the current mammography requirements. Images can be reviewed on the Hologic's SecureView DX Diagnostic Workstation (K062107) and with hard copy printers that are approved for mammography.

D. Radiation Dose

The Mammography Quality Standards Act (MQSA) states: "The average glandular dose delivered during a *single* craniocaudal view of an FDA-accepted phantom simulating a standard breast shall not exceed 3.0 milligray (mGy) (0.3 rad) *per exposure*."[21 CFR 900.12(e)(5)(vi)] The total dose for both the 2D FFDM and 3D DBT was measured using an ACR breast phantom that simulates a 4.2 cm thick, 50% fat / 50% glandular equivalent compressed breast. The total dose to the phantom for the 2D plus 3D (4-views total) is approximately 5 mGy. The total dose to the phantom for the 2D plus 3D MLO (3-views total) in Reader Study 2 is approximately 3.5 mGy.

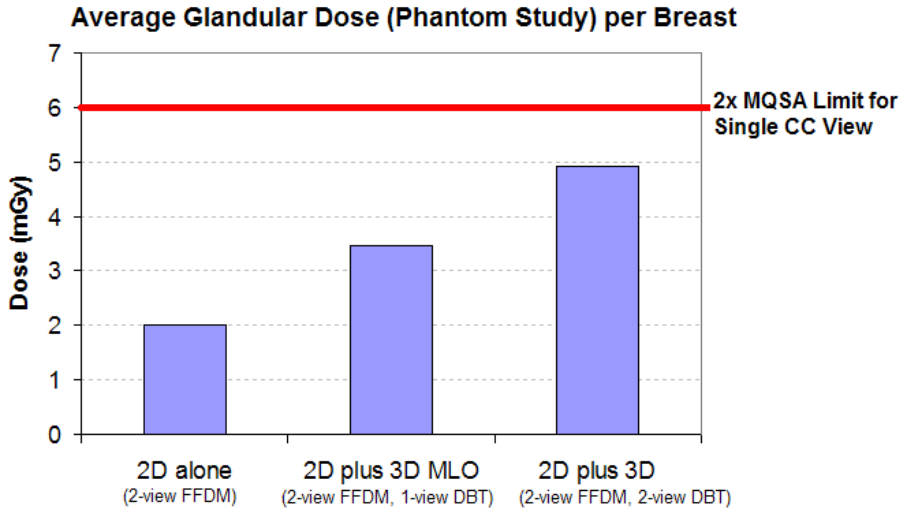


Figure 1: Comparison of the average glandular dose (per breast) from the three modes. The glandular dose to a phantom simulating a standard breast for 2D plus 3D would be less than twice the MQSA limit for single CC view.

III. Reader Studies

Hologic submitted Reader Study 1 as the pivotal study of the PMA. Hologic submitted Reader Study 2 and the study from the University of Pittsburgh in response to deficiencies from FDA. In particular, the results of Reader Study 2 were provided to support a lower dose tomosynthesis protocol (i.e., 2D plus 3D MLO) as well as address concerns with the reader scoring methodology (i.e., patient-based vs. lesion-based) and impact of 2D plus 3D on the recall rate of cancer patients.

Note: The panel binder includes the protocols for both Reader Study 1 and Reader Study 2.

A. Reader Study 1 (Pivotal Study)

i. Study Design

OBJECTIVE	To compare the clinical performance of conventional 2D FFDM plus 3D tomosynthesis images to those of 2D FFDM images alone.
PATIENT POPULATION	1192 Subjects from 5 clinical US centers; subjects were from either a screening group or a biopsy group.
IMAGING	2-view 2D FFDM (MLO and CC) 2-view 3D DBT (MLO and CC)
GROUND TRUTH	Two different radiologists classified each image to one of the following four categories: 1. Negatives: Scored BIRADS 1 or 2 by both the 2D site reader and the 3D site reader; truth was based on a negative independent double reading. 2. Recalls: Screening cases recalled (BIRADS=0) by either the 2D or 3D site reader. Each patient was followed to determine the outcome of the additional imaging or biopsy. Truth was determined from the imaging results unless a biopsy was performed. If a biopsy was done, the pathology results were used to determine the truth. 3. Benign: Biopsy cases where pathology/histology demonstrated a benign lesion. 4. Cancers: Biopsy cases where pathology/histology demonstrated a cancer.
STUDY DATA	75 Negatives 141 Recalls (does not include cancer patients) 48 Benign 48 Cancers Of the original 1192 subjects, 1083 subjects were eligible to be randomized into the reader study. 316 were randomized into the study, but 4 cases were excluded for the presence of radio opaque markers on the images found during the image review. The final dataset consisted of 312/1083 (28.8%) subjects, 222/856 (25.9%) subjects from the Screening Group and 90/227 (39.7%) subjects from the Biopsy Group. All pathology proven cancers that had complete data and passed the quality control review were included. Negative, benign and recall (non-cancer) cases were randomly selected.
READERS	14 Radiologists participated in the reader study. Only the results of the 12 radiologists who successfully completed the reader training were included in the analysis. The 12 readers, included: 5 highly experienced, 2 experienced, and 5 less experienced radiologists.
IMAGE SCORING	Readers were asked to provide the following information after reading the 2D images and then after reading both the 2D plus 3D images. 1. Probability of Malignancy (POM) score of 0 to 100 *(for ROC analysis)*; 2. Initial BIRADS score of 0,1, or 2 *(for recall analysis)*; and 3. Forced BIRADS score of 1, 2, 3, 4, or 5 when initial BIRADS score is 0 *(for sensitivity/specificity analysis)*

PRIMARY ENDPOINTS	(a) 0.05 increase in the ROC area under the curve (AUC) (b) Reduction of 20% in recall rate Comparison: 2D FFDM alone verses 2D FFDM plus 3D DBT

The study was designed to have a minimum of 10 readers reading all cases, a minimum of 75 recall screening cases, and a minimum of 50 cancer cases. The study included fourteen readers, but only the results of the 12 radiologists who successfully completed all of the reader training requirements were included in the analysis.

The reader scores were used to perform three main analyses:

1) ROC area under the curve (AUC), based on the probability of malignancy (POM);
2) Recall, based on the reader scores of BIRADS (0,1,2); and
3) Sensitivity and specificity based on the "Forced BIRADS" scores.

The primary study objectives were to show a 20% decrease in recall rate among the screening cases and a 0.05 increase in area under the ROC curve (AUC) using all cases either based on a probability of malignancy scale (1-100) or a forced BIRADs scale (initial BIRADS 0 score in recall analysis forced to BIRADs 1-5). After the study was completed and the PMA submitted, the study objectives were clarified in a subsequent meeting with FDA on August 25th, 2008, and in the sponsor's responses to FDA deficiencies. The success criteria were actually to show any statistically significant improvement in AUC and any statistically significant improvement in recall rate among the screening cases.

ii. Study Population

The pivotal study (Reader Study 1) enrolled a subset of screening and diagnostic cases from larger cohorts of these two groups having a total of 1192 cases; 109 subjects were discontinued, leaving 1083 for possible enrollment into the study.

Data from 316 subjects was randomized into Reader Study 1, but 4 cases were excluded for the presence of radio opaque markers on the images found during the image review. The final dataset consisted of 312/1083 (28.8%) subjects, 222/856 (25.9%) subjects from the Screening Group and 90/227 (39.7%) subjects from the Biopsy Group. Reader Study 1 contained both recall cases from the screening group and cancer cases found among women scheduled for biopsy based on 2D imaging.

Table 2 presents the distribution of calcification and non-calcification in the dataset. The distribution of breast density (fatty or dense) of the cases in the study was not available at the time of writing.

Table 2: Distribution of calcification and non-calcification for 308 of the 312 cases in the reader study.

Study	Case Type	Calcification	Non-Calcification	Total
Reader Study 1	Cancer	24	24	48
	Recall	31	108	139
	Benign	28	19	47
	Negative	0	74	74
	Total	83	225	308

The mean age of all subjects in the study (n=316) was 54.5 (range 25.0-88.0); the mean ages were 56.8, 51.7, 54.6, and 54.8 for cancer, benign, recall, and negative subjects, respectively. Correspondingly, the mean ages for subjects not in the study (n=632) 53.7, 50.1, 53.2, and 54.2 for cancer, benign, recall, and negative subjects, respectively.

Fourteen of the 62 cancers were excluded from the dataset. Four were excluded because final pathology report was unavailable and the subject was not monitored at the time of reader study randomization. Four were excluded because of incomplete 2D data, three because of motion artifact, and one each because of prior clip placement or surgery, incomplete 3D data, and BB marker over the area of interest. A total of 48 cancer cases were used in Reader Study 1, with 5 of the cancer cases coming from the Screening Group and 43 of the cancer cases coming from the Biopsy Group; 16 of the cancers DCIS only and 32 of the cancers were invasive.

Table 3: Mean and median sizes (mm) for 48 cancers in Reader Study 1.

Summary	DCIS Max Dimension	DCIS Mean Dimension	Invasive Cancer Max Dimension	Invasive Cancer Mean Dimension
Mean size (mm)	15.6	14.0	16.5	15.6
Median size (mm)	10.0	9.8	14.5	14.0

iii. Results

Scores were compiled for each reader for each of the 312 cases. For the recall analysis, all 312 cases were analyzed. For the ROC analysis, five cases were found to be missing a score for 1 of the 12 readers. One of these five cases was missing a single forced BIRADS score; the others were missing both a forced BIRADS score and a probability of malignancy score. Since the ROC program being used required all data for all cases, 4 cases were not used in the POM ROC analysis (2 recall cases, 1 benign and 1 negative) and 5 cases were not used in the FB ROC analysis (2 recall cases, 2 benign and 1 negative). Therefore, there were 308 cases used for the POM ROC analysis and 307 cases used for the forced BIRADS ROC analysis.

ROC Area under the curve (AUC)

For all readers the ROC AUC was superior with 2D plus 3D imaging compared with 2D alone. The average increase was 0.072 (95% CI: 0.037 to 0.107) using the probability of malignancy scale (p-value 0.0001) and 0.071 (95% CI 0.034 to 0.110) using the forced BIRADs scale (p-value 0.0004).

The p-values were calculated using the program DBM MRMC 2.1 Beta 2 developed by Hillis, Schartz, and Bernbaum (http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index6.htm). The program uses a modification of the jackknife pseudovalue approach by Dorfman, Bernbaum, and Metz (*Invest-Radiol* 1992; 27:723-731, as described in HIllis, Bernbaum, and Metz (Acad Radiol 2008; 15:647–661).

The sponsor also constructed smoothed ROC curves for both modalities by pooling reader scores. The curve was fitted using the program PROPROC (also available at the same URL given above). The program fits a smoothed ROC curve by maximum likelihood using a "proper" binormal model. FDA notes that pooling reader scores can attenuate the ROC curve relative to the per reader ROC curves. The degree of attenuation can vary by modality. However, our understanding is that these smoothed ROC curves were not used to establish statistical significance.

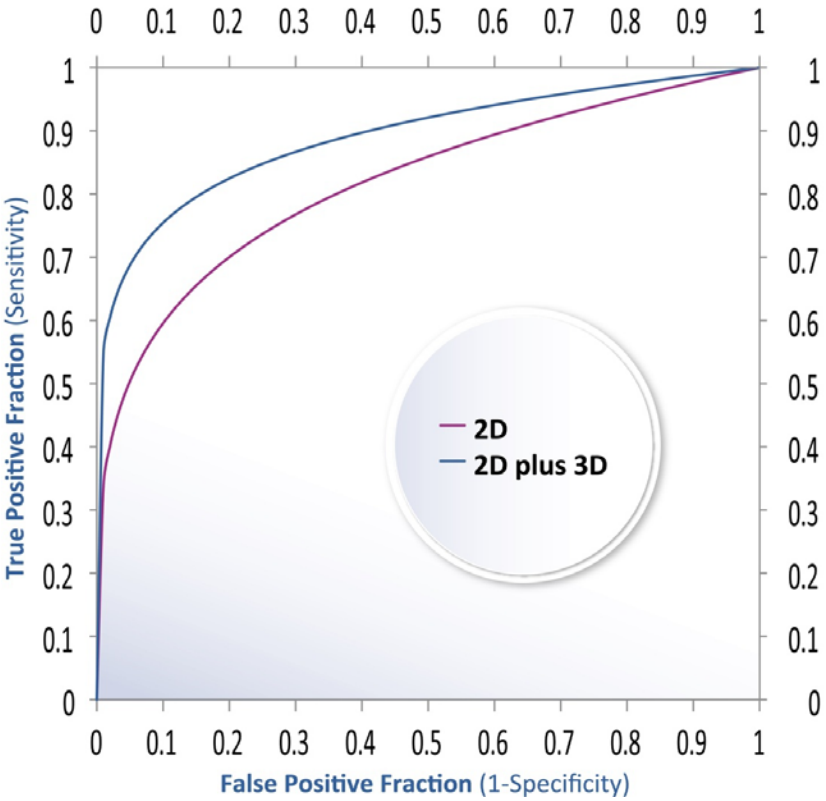


Figure 2: (Reader Study 1) ROC curves comparing reader performance with 2D alone to 2D plus 3D. 95% confidence intervals for the change in ROC AUC are presented in the text.

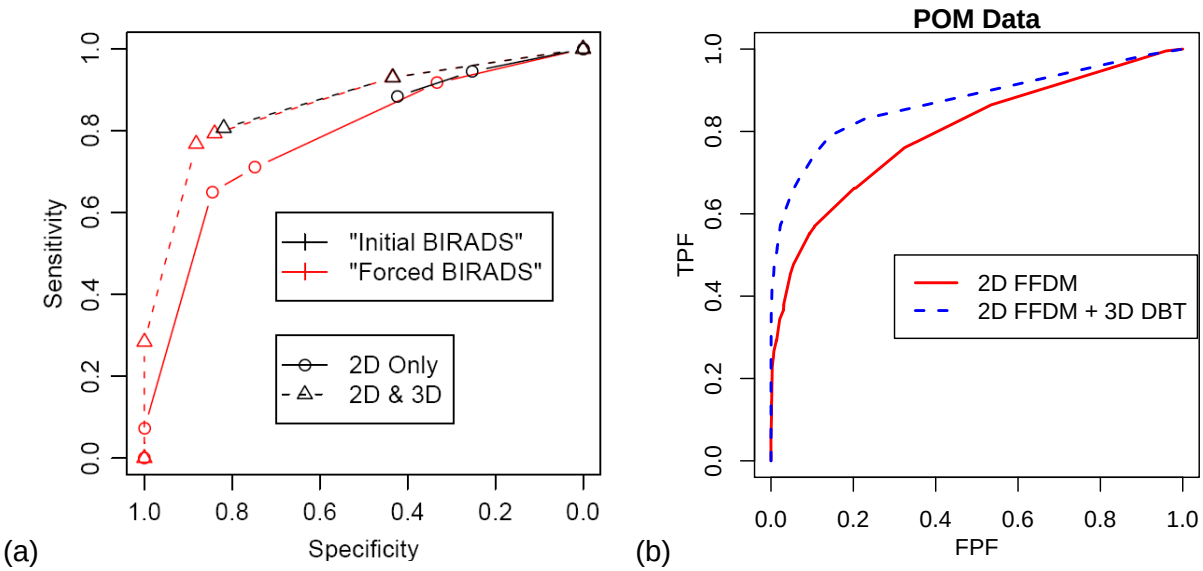


Figure 3: (Reader Study 1) As an alternative to the smoothed ROC curves given by the sponsor and displayed above, FDA is providing the empirical ROC curves comparing reader performance with 2D alone to 2D plus 3D based on (a) BIRADS scores and (b) probability of malignancy (POM).

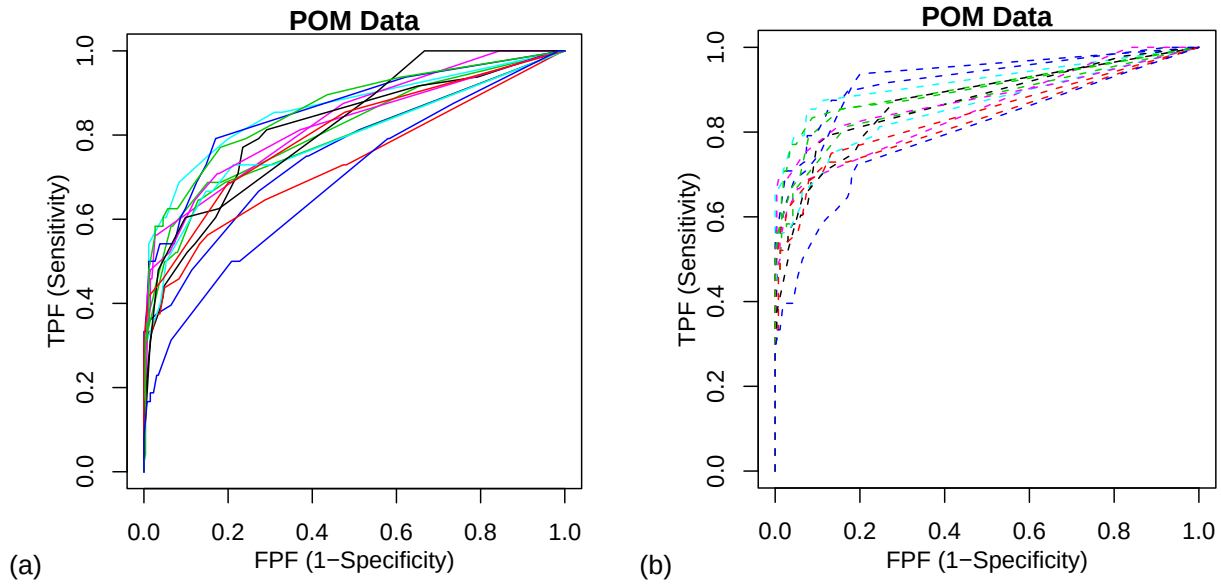
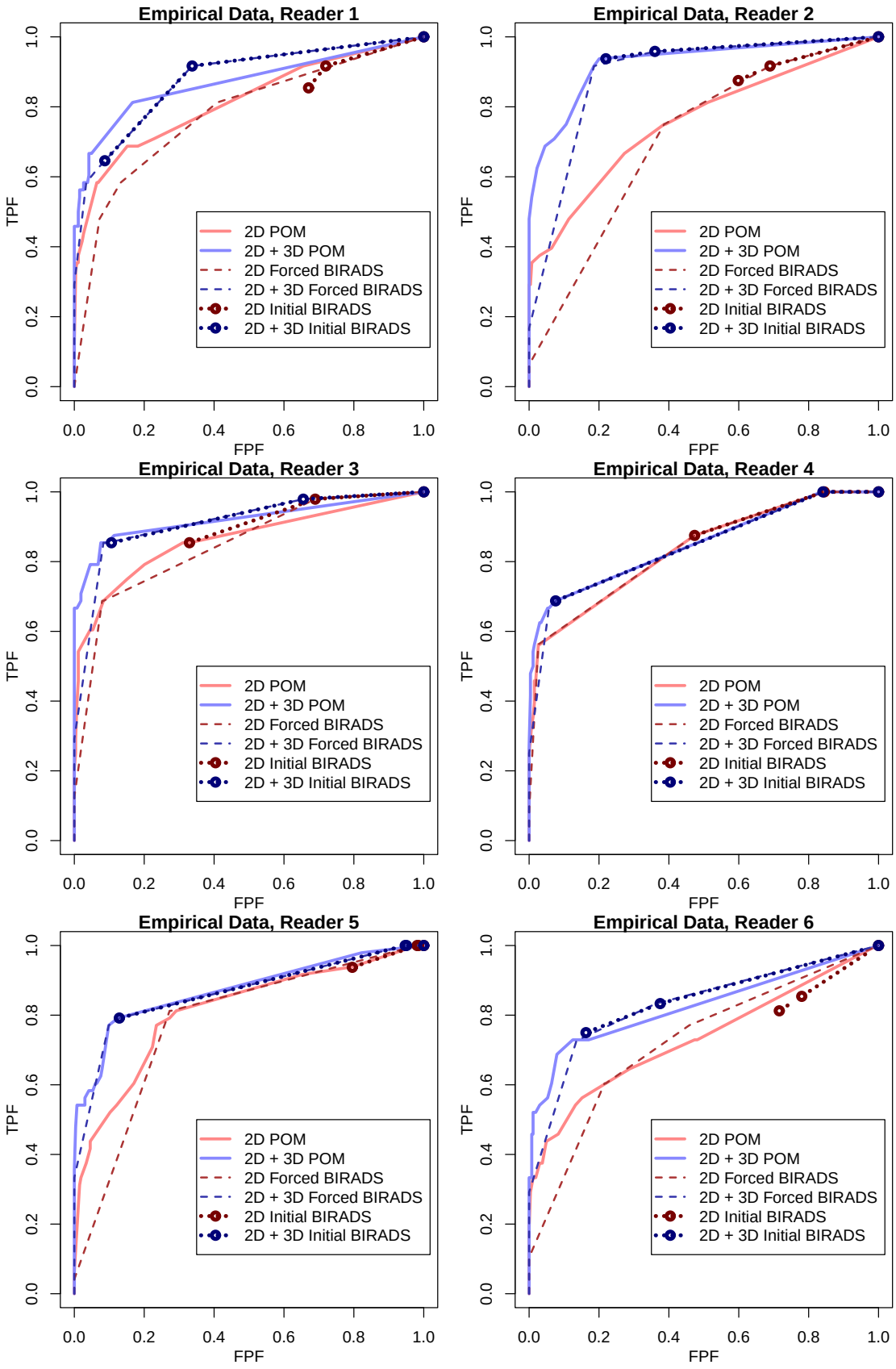
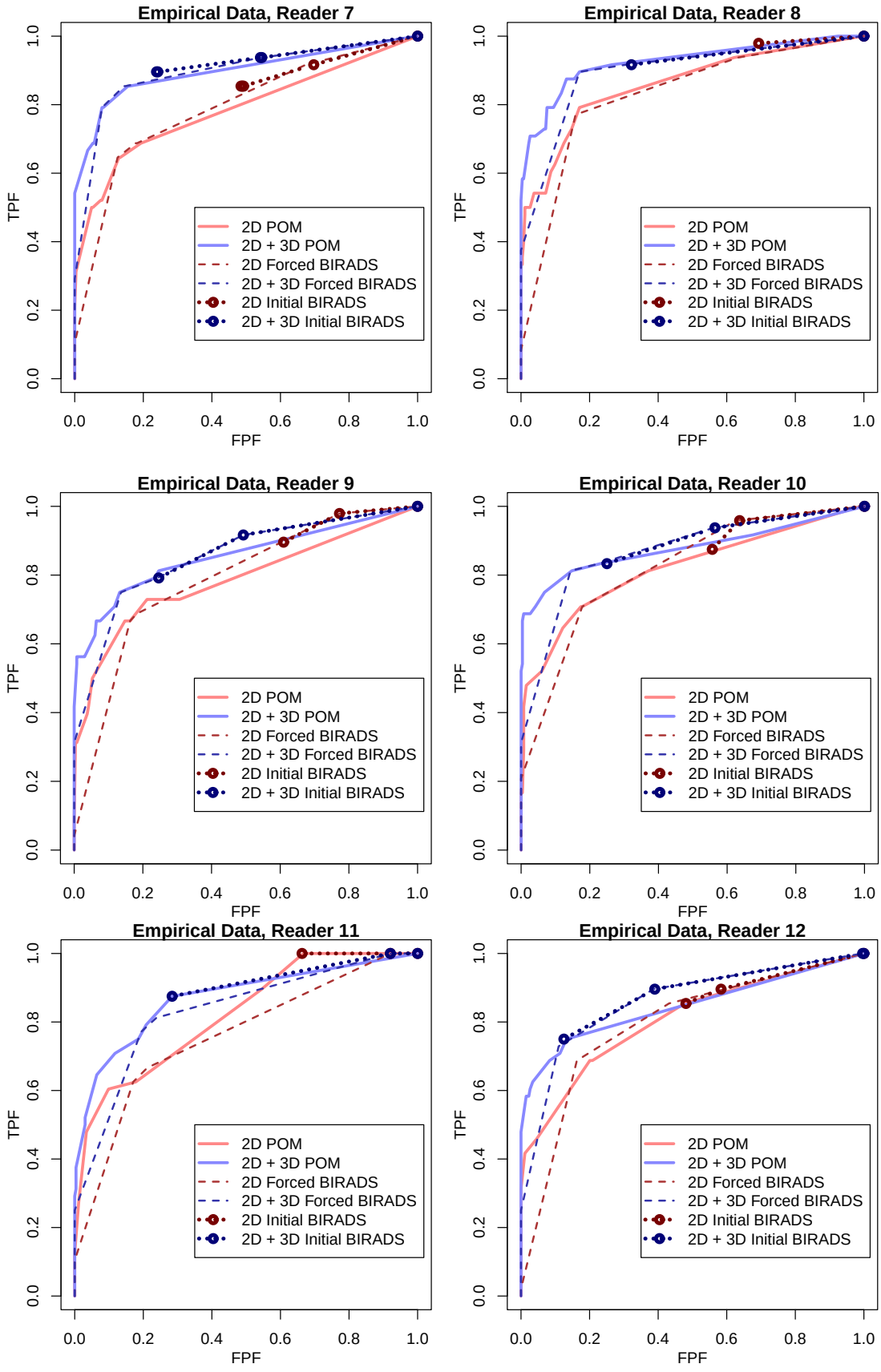
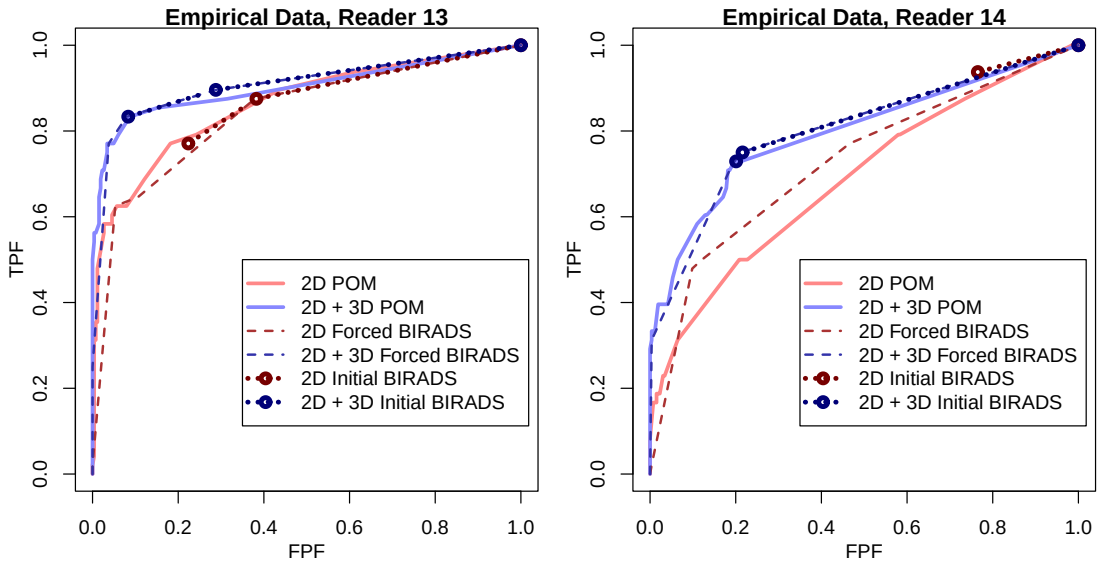


Figure 4: (Reader Study 1) FDA is providing the empirical ROC curves based on probability of malignancy (POM) for each reader when using: (a) 2D alone and (b) 2D plus 3D.

In the following plots, the empirical ROC curves are displayed by reader for the probability of malignancy, forced BIRADS, and initial BIRADs scales. The two excluded readers (8 and 14) from Reader Study 1 are also presented for completeness.


Empirical Data, Reader 1
Empirical Data, Reader 2
Empirical Data, Reader 3
Empirical Data, Reader 4
Empirical Data, Reader 5
Empirical Data, Reader 6
TPF
FPF
0.0
0.2
0.4
0.6
0.8
1.0
2D POM
2D + 3D POM
2D Forced BIRADS
2D + 3D Forced BIRADS
2D Initial BIRADS
2D + 3D Initial BIRADS


Empirical Data, Reader 7
Empirical Data, Reader 8
Empirical Data, Reader 9
Empirical Data, Reader 10
Empirical Data, Reader 11
Empirical Data, Reader 12
TPF
FPF
0.0
0.2
0.4
0.6
0.8
1.0
2D POM
2D + 3D POM
2D Forced BIRADS
2D + 3D Forced BIRADS
2D Initial BIRADS
2D + 3D Initial BIRADS



Recall Rate

The primary objective for recall rate was to show a significant reduction in the recall rate among the screening cases. For screening cases the recall rate went from 51.5% with 2D to 12.9% with 2D plus 3D. This reduction was significant for all readers. The recall rate was also calculated among all cases in both the screening and biopsy groups as well just in the screening group. The recall rate was also stratified by cancer versus. non-cancer cases.

There was a significant reduction (p-value < 0.05) in the recall rate (score BIRADS=0) for non-cancer patients for all 12 readers when using 2D plus 3D compared to 2D alone (N=264). However, there was also a significant reduction in the recall rate for cancer patients for 4 of the 12 readers. The average decrease was from 87.2% (2D alone) to 80.4% when using 2D plus 3D (N=48). For the cancer patients, the recall rate increased for three of the twelve readers, stayed the same for one reader, and non-significantly decreased for the remaining four readers. The power to detect significant differences for readers is limited by the small sample size of 48 cancer patients.

Table 4: Recall rates for Reader Study 1 (averaged over all readers); 95% confidence intervals calculated by FDA.

Study	Mode	Cancer Recall Rate (N=48)	Non-Cancer Recall Rate (N=264)	Screening Cases, Recall Rate†
Reader Study 1	2D	87.2% (81%, 94%)	55.1% (48%, 66%)	51.5% (41%, 61%)
	2D plus 3D	80.4% (71%, 89%)	16.7% (13%, 23%)	12.9% (8.9%, 17%)

†Primary analysis objective

Forced BIRADS Analysis (Sensitivity and Specificity)

Estimates of sensitivity and specificity were calculated based on "Forced BIRADS" scores using BIRADS 4 as the cut-off. When taking BIRADS 4 or 5 as positive, the sensitivity increases 10.7% and specificity increases 5.1% with comparing using 2D plus 3D with using 2D alone.

Table 5: Sensitivity and specificity for "Forced BIRADS" with BIRADS 4 as the cutoff (averaged over readers); 95% confidence intervals calculated by FDA.

	Mode	Sensitivity (N=48)	Specificity (N=264)
Reader Study 1	2D	65.5% (54%, 78%)	84.1% (80%, 90%)
	2D plus 3D	76.2% (67%, 85%)	89.2% (84%, 92%)

When sensitivity and specificity are calculated with BIRADS 3, 4, or 5 as positive, the sensitivity increases 8.0% and specificity increases 8.7% with the addition of the 3D tomosynthesis images.

Table 6: Sensitivity and specificity for "Forced BIRADS" with BIRADS 3 as the cutoff (averaged over readers).

	Mode	Sensitivity (N=48)	Specificity (N=264)
Reader Study 1	2D	70.8%	76.9%
	2D plus 3D	78.8%	85.6%

Calcifications and Non-calcifications

ROC analysis was stratified by calcifications and non-calcifications. The difference in AUC between 2D FFDM and 2D plus 3D DBT was 10.1% for non-calcifications and statistically significant (p-value 0.0008), but was only 3.5% for calcifications and not statistically significant (p-value 0.073). Besides the smaller difference, the smaller sample size for calcifications may have contributed to the lack of statistical significance.

Table 7: ROC AUC stratified by Calcification Status.

	Statistic	All Cases	Calcifications	Non-Calcifications
Reader Study 1	2D AUC	83.0%	80.4%	81.7%
	2D plus 3D AUC	90.2%	84.0%	91.8%
	Difference	7.1%	3.5%	10.1%
	p-value	0.0004	0.073	0.0008
	95% CI	3.3, 11.0%	–0.4, 7.4%	4.3, 15.9%

The sensitivity and specificity using BIRADS 4 or 3 as the cut-off were also stratified by calcifications and non-calcifications. The increases in sensitivity and specificity due to 3D DBT are much larger for non-calcifications than for calcifications, whether the BIRADs cut-off was 3 or 4. Statistical significance of the differences was not assessed

Table 8: Sensitivity and Specificity by Calcification Status, BIRADS 4 as cut-off (per protocol).

	Statistic	All Cases	Calcifications	Non-Calcifications
Reader Study 1	2D Sensitivity	65.5%	72.9%	58.0%
	2D plus 3D Sensitivity	76.2%	75.0%	77.4%
	Difference	10.7%	2.1%	19.4%
	2D Specificity	84.1%	71.6%	87.7%
	2D plus 3D Specificity	89.2%	74.3%	93.5%
	Difference	5.1%	2.7%	5.7%

Table 9: Sensitivity and Specificity by Calcification Status, BIRADS 3 as cut-off.

	Statistic	All Cases	Calcifications	Non-Calcifications
Reader Study 1	2D Sensitivity	70.8%	78.5%	63.2%
	2D plus 3D Sensitivity	78.8%	79.5%	78.1%
	Difference	8.0%	1.0%	14.9%
	2D Specificity	76.9%	61.0%	81.4%
	2D plus 3D Specificity	85.6%	66.5%	91.0%
	Difference	8.7%	5.5%	9.6%

Study Recall Rate When Adjusted to Entire Screening Cohort

Among negatives and recalls in the screening group, the study recall rates were 0.515 for 2D and 0.129 for 2D plus 3D. These recall rates are high because the study was enriched for recalls relative negatives, as the following table shows:

Table 10: Breakdown of the negative and recalls in the screening group and the recall rate in the reader study.

Study	Case Type	In Reader Study	Not In Reader Study	Total	Reader Study Analysis Set	Reader Study Recall Rate	
						2D	2D plus 3D
Reader Study 1	Negative	75	490	565	74	0.327	0.050
	Recall	143	134	277	141	0.615	0.171
	Total	218	624	842	215	0.515	0.129

The recall rates can be adjusted by considering the recall rates by case type (negative, recall) and applying these rates to the fractions of these types that were observed in the entire cohort before random selection into the study. The adjusted recall rates are then 0.422 for 2D and 0.090 for 2D plus 3D. These can be compared with the fraction of recalls in the cohort, 0.329 (277/842). Before the study, case type (recalls, negatives) were defined by assimilating 2D and 3D evaluations of each case made independently by separate readers.

Ground Truth Determination

Many of the confirmed cancer cases had a conventional 2D FFDM examination that resulted in a recall recommendation prior to the 3D tomosynthesis procedure. Therefore, the study may have been biased in favor of the 2D FFDM mode because the selection cancer cases may have excluded cases that might have been recalled with the use of 3D tomosynthesis.

In addition, cases considered negative for cancer were not followed for confirmation. Therefore, some cancer cases may have been misclassified as negative according to the operational definition of ground truth used in the study. However, the likelihood of this may have been low given the low prevalence of cancer. Use an estimate of 0.005 as the cancer rate among screening patients, then less than 1 of the 75 negative screening cases used for the reader study would be expected to have cancer (75*0.005 = 0.375).

Risk / Benefit

A risk verses benefit analysis for screening and diagnostic populations can be constructed from the results of the study and assumed prevalences of cancer for these populations. The sponsor conducted their own risk / benefit analysis, which was based on a number of assumptions from the study, cancer prevalence, and radiation risk. However, the study enrichment makes it difficult to generalize the estimates of recall, sensitivity, specificity and any risk / benefit analysis based on these measures.

Positive and negative predictive values (PPV and NPV) are not presented because they depend on the prevalence of cancer, which was highly enriched in the study relative to the prevalences seen in screening or diagnostic work-up populations. However, positive and negative likelihood ratios do not depend on prevalence, and improved positive and negative likelihood ratios imply improved PPV and NPV, respectively (Biggerstaff, *Stat Med*, 2001).

Using the recall sensitivities and the adjusted recall specificities for Reader Study 1, the positive likelihood ratio was 2.07 for 2D and 8.93 for 2D plus 3D, indicating that PPV was greater for 2D plus 3D for screening. The negative likelihood ratio was 0.222 for 2D and 0.215 for 2D plus 3D, indicating that NPV was about the same for both modes. Statistical significance was not assessed.

Using the sensitivities and specificities with forced BIRADS 4 as the cut-off, the positive likelihood ratio was 1.58 for 2D and 4.81 for 2D plus 3D, indicating that PPV was greater for 2D plus 3D for diagnostic purposes. The negative likelihood ratio was 0.285 for 2D and 0.235 for 2D plus 3D, indicating that NPV was also greater for 2D plus 3D than for 2D alone. Statistical significance was not assessed.

B. Reader Study 2

Hologic submitted Reader Study 2 in response to deficiencies from FDA. In particular, the results of Reader Study 2 were provided to support a lower dose tomosynthesis protocol (i.e., 2D plus 3D MLO) as well as address concerns with the reader scoring methodology (i.e., identification of the correct location of a malignant lesion was not required for crediting readers with a true positive result) and impact of 2D plus 3D on the recall rate of cancer patients. Reader Study 2 used new readers and a new random selection of non-cancer cases. Reader Study 2 reused the cancer cases from Reader Study 1 with the addition of three more cancers.

i. Study Design

OBJECTIVE	*To investigate the potential of an alternative, lower dose option (2D plus 3D MLO); to include scoring for a lesion-based analysis.*
PATIENT POPULATION *(same as Reader Study 1)*	1192 Subjects from 5 clinical US centers; subjects were from either a screening group or a biopsy group.
IMAGING *(same as Reader Study 1)*	2-view 2D FFDM (MLO and CC) 2-view 3D DBT (MLO and CC)
GROUND TRUTH *(same as Reader Study 1)*	Two different radiologists classified each image to one of the following four categories: 1. Negatives: Scored BIRADS 1 or 2 by both the 2D site reader and the 3D site reader; truth was based on a negative independent double reading. 2. Recalls: Screening cases recalled (BIRADS=0) by either the 2D or 3D site reader. Each patient was followed to determine the outcome of the additional imaging or biopsy. Truth was determined from the imaging results unless a biopsy was performed. If a biopsy was done, the pathology results were used to determine the truth. 3. Benign: Biopsy cases where pathology/histology demonstrated a benign lesion. 4. Cancers: Biopsy cases where pathology/histology demonstrated a cancer.
STUDY DATA *(3 new cancers; new random selection of non-cancers cases)*	74 Negatives 138 Recalls (does not include cancer patients) 47 Benign 51 Cancers The 48 cancer cases used in Reader Study 1 were supplemented with 3 cancer cases for which the pathology reports had not previously been available. A new randomized selection of non-cancer cases was performed; 123/259 (47%) of the non-cancer cases were the same in Reader Study 1 and Reader Study 2.
READERS *(new radiologists, additional training)*	15 new radiologists (5 highly experienced, 6 experienced, and 4 less experienced). Readers received additional 3D training based on the types of errors made during Reader Study 1.
IMAGE SCORING *(addition of lesion based scoring, #4)*	Readers were asked to provide the following information after reviewing a full set of 2D images (CC and MLO), followed by the 2D plus 3D MLO images, followed by the full set of 2D plus 3D images. 1. Probability of Malignancy (POM) score of 0 to 100 *(for ROC analysis)*; 2. Initial BIRADS score of 0,1, or 2 *(for recall analysis)*; and 3. Forced BIRADS score of 1, 2, 3, 4, or 5 *(for sensitivity/specificity analysis)* *4. Breast location and lesion type (calcification, mass/architectural distortion or asymmetry) for any lesion that would result in a recall (initial BIRADS 0)*

PRIMARY ENDPOINTS *(new 2D plus 3D MLO comparisons)*	(a) Improved ROC area under the curve (AUC) and/or (b) Reduced recall rate Comparisons: (a) 2D versus 2D plus 3D *(b) 2D versus 2D plus 3D MLO* *(c) 2D plus 3D verses 2D plus 3D MLO*

The 2D and 3D images used in Reader Study 2 were obtained from the same set of images acquired for the Reader Study 1. From the 1083 eligible subject images for randomization, 310 images were selected and randomized into Reader Study 2; 220/1083 (20.3%) subjects were from the Screening Group and 90/1083 (8.3%) subjects were from the Biopsy Group, for a total of 28.6% of the images from the acquisition study. The 48 cancer cases used in Reader Study 1 were supplemented with 3 cancer cases for which the pathology reports had not previously been available. For non-cancer cases, a new randomized selection of cases was performed; 123/259 (47%) of the non-cancer cases were the same in Reader Study 1 and Reader Study 2.

The power analysis performed by an independent statistician determined the sample size requirements of this study based on having 10 or more radiologists score the images. Fifteen radiologists participated in the reader study to allow for potential scheduling conflicts or failure to pass the training evaluation test.

The training of the radiologists for Reader Study 2 was influenced by the types of reading errors made during the previous study. In Reader Study 1, the readers were trained not to dismiss lobulated masses even if they were circumscribed; however, based upon review of the dismissed cancer cases approximately half of the readers did not adhere to that training. In Reader Study 2, the readers were again trained not to dismiss lobulated circumscribed masses and their training was reinforced in written format and with further examples.

ii. Results

The sponsor performed the same three types of analyses as in Reader Study 1 (i.e., ROC, recall, and forced BIRADS). For the recall analysis, all 310 cases were analyzed; however, for the ROC analysis, seven cases were found to be missing a score due to operator errors and were omitted from the probability of malignancy ROC analysis and BIRADS ROC analysis. To aid the discussion, the results from Reader Study 1 and Reader Study 2 are shown together.

ROC Area under the curve (AUC)

The best reader performance was achieved when using 2D plus 3D (all views). The lower dose 2D plus 3D MLO option was also superior to 2D alone. The following are the results using the probability of malignancy scores; the BIRADS based change in ROC AUC is similar.

- The 2D plus 3D mode was superior to 2D alone; ROC AUC improved by 0.068 (95% CI 0.041 to 0.095) with a p-value < 0.0001. All 15 radiologists had a higher AUC using 2D plus 3D compared to 2D images alone.

- The 2D plus 3D MLO mode was superior to 2D alone; ROC AUC improved by 0.036 (95% CI 0.009 to 0.063) with a p-value of 0.0087. All 15 radiologists had a higher AUC using 2D plus 3D MLO compared to 2D alone.

- The 2D plus 3D mode was superior to 2D plus 3D MLO; ROC AUC improved by 0.032 (95% CI 0.005 to 0.059) with a p-value of 0.0205. 13 of 15 radiologists had a higher AUC using 2D plus 3D compared to 2D plus 3D MLO.

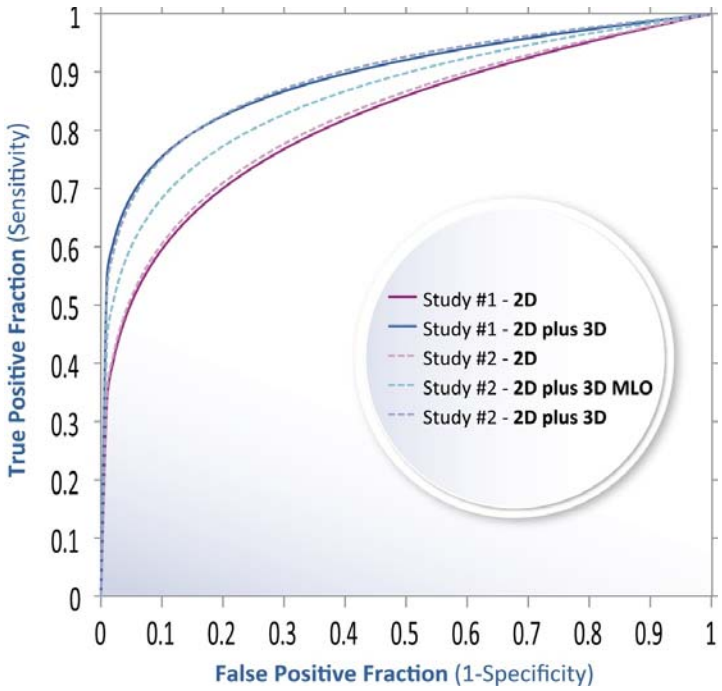


Figure 5: ROC curves for Reader Study 1 and Reader Study 2. In both studies, the ROC curves for the 2D alone and the 2D plus 3D modes are essentially the same. In Reader Study 2, the ROC curve for 2D plus 3D MLO is between the 2D alone and 2D plus 3D modes. 95% confidence intervals for the change in ROC AUC are presented in the text.

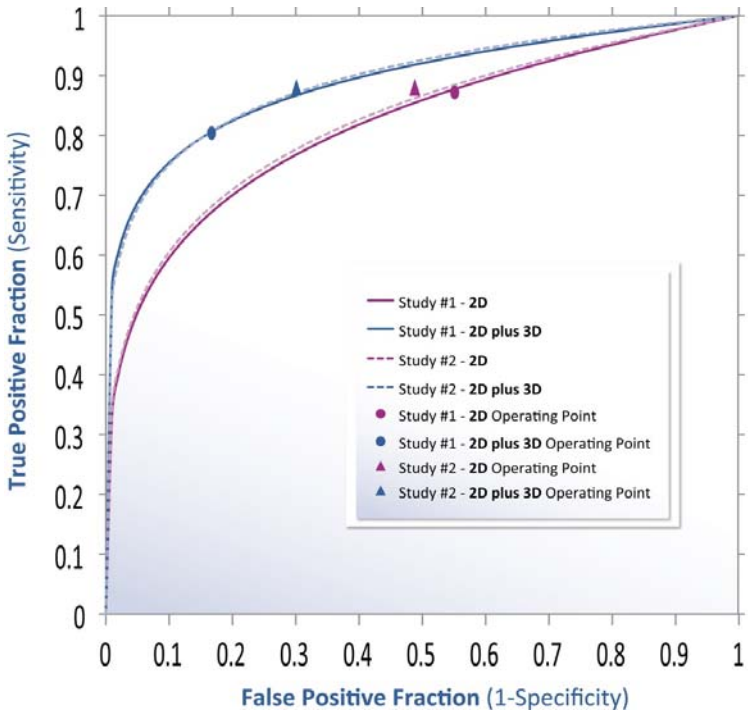


Figure 6: ROC curves and operating points from Reader Study 1 and Reader Study 2. The ROC curve is parametric fit based on the probability of malignancy scoring. The operating points are based on the recall rate. The ROC curves are essentially unchanged, but there is a shift in the reader's operating points.

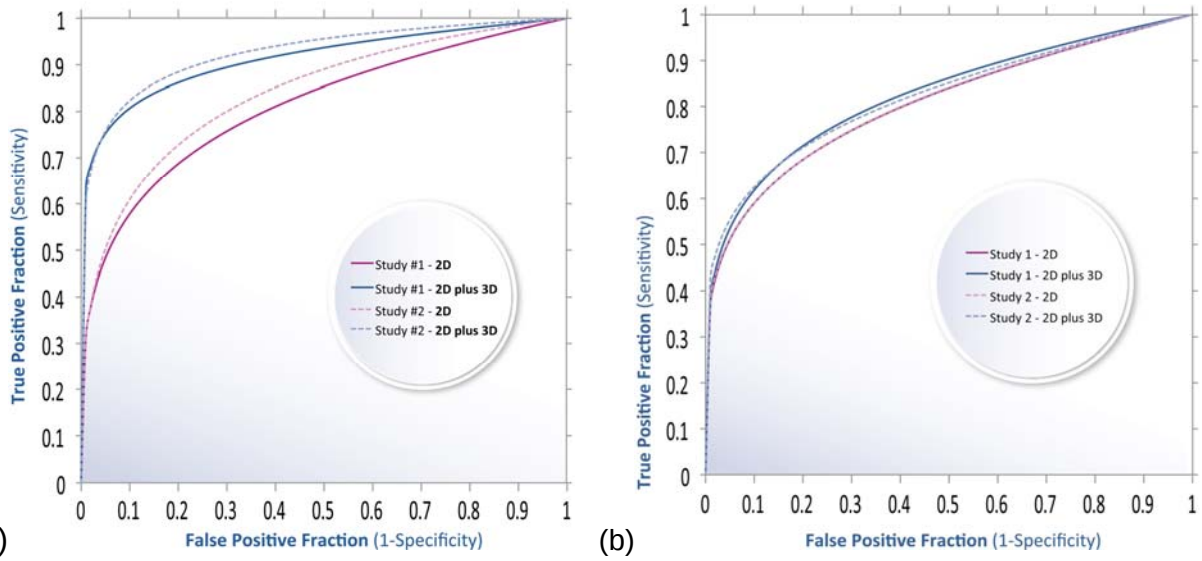


Figure 7: ROC curves from Reader Study 1 and Reader Study 2 for (a) non-calcification and (b) calcifications. The gains in ROC AUC for calcifications were not significant while those for non-calcification cases were highly significant in both Reader Study 1 and Reader Study 2.

Recall Rate

The screening recall rates were 44.2% for 2D alone, 27.2% for 2D plus 3D MLO and 24.0% for 2D plus 3D. The differences in recall rates between 2D plus 3D MLO and 2D alone were significant for all readers. The differences in recall rates between 2D plus 3D and 2D alone were significant for all readers.

There was a significant reduction in the recall rate (score BIRADS=0) for non-cancer patients when using 2D plus 3D (or 2D plus 3D MLO) compared to 2D alone. In Reader Study 2, the recall rate for cancer patients was 88.0% for both 2D plus 3D and 2D alone; for the 2D plus 3D MLO, the recall rate for cancer patients was lower at 84.2%. The results were similar with lesion-based scoring (i.e., when correct lesion type and location were required).

Table 11: Recall rate for Reader Study 1 and Reader Study 2; additional information to evaluate lesion-based scoring was collected in Reader Study 2.

Study	Scoring Method	Mode	Cancer Recall Rate	Non-Cancer Recall Rate	Screening Cases, Recall Rate
Reader Study 1	Patient-based	2D	87.2%	55.1%	51.5%
		2D plus 3D	80.4%	16.7%	12.9%
Reader Study 2	Patient-based	2D	88.0%	48.8%	44.2%
		2D plus 3D MLO	84.2%	32.7%	27.2%
		2D plus 3D	88.0%	30.1%	24.0%
	Lesion-based	2D	84.8%	48.8%	*
		2D plus 3D MLO	81.5%	32.7%	*
		2D plus 3D	85.7%	30.1%	*

*Not available at time of writing.

Forced BIRADS Analysis (Sensitivity and Specificity)

Estimates of sensitivity and specificity were calculated based on the forced BIRADS scores. For Reader Study 2, when taking BIRADS 4 or 5 as positive, the sensitivity increases 16.0% and specificity decreases 1.7% with comparing using 2D plus 3D with using 2D alone.

Table 12: Sensitivity and specificity from the "Forced BIRADS" (using BIRADS scores 4 and 5 as positive).

	Mode	Sensitivity (N=48)	Specificity (N=264)
Reader Study 1	2D	65.5%	84.1%
	2D plus 3D	76.2%	89.2%
Reader Study 2	2D	62.7%	86.2%
	2D plus 3D MLO	71.4%	86.0%
	2D plus 3D	78.7%	84.5%

For Reader Study 2, when taking BIRADS 3, 4, or 5 as positive, the sensitivity increases 12.0% and specificity increases 1.7% with comparing using 2D plus 3D with using 2D alone.

Table 13: Sensitivity and specificity from the "Forced BIRADS" (using BIRADS scores 3, 4, and 5 as positive).

	Mode	Sensitivity (N=48)	Specificity (N=264)
Reader Study 1	2D	70.8%	76.9%
	2D plus 3D	78.8%	85.6%
Reader Study 2	2D	70.5%	78.3%
	2D plus 3D MLO	75.0%	80.8%
	2D plus 3D	82.5%	80.0%

C. Pittsburgh Reader Study

Hologic also submitted a research study performed by the University of Pittsburgh.[1] The research study is relevant to the discussion, but the study was not designed or intended for the FDA review in this PMA submission. The study used different cancer cases than Reader Study 1 and Reader Study 2.

Eight readers read 125 cases in each of four display conditions: 2-view FFDM alone, 11 low-dose projections, reconstructed digital breast tomosynthesis images (MLO and CC), and a combined display mode of FFDM and digital breast tomosynthesis images. The cases were read in four reading sessions separated by at least one month. A single mode was used in each reading session. Two readers initially read by FFDM alone, two readers began with the frames, two readers began with the digital breast tomosynthesis images, and two readers began with the combined FFDM and tomosynthesis display. Evidently, a latin square design was used in modality, reader pairs, and reading session. The design should eliminate confounding of modality effects with reading session effects.

i. Abstract

OBJECTIVE. The purpose of this study was to compare in a retrospective observer study the diagnostic performance of full-field digital mammography (FFDM) with that of digital breast tomosynthesis.

MATERIALS AND METHODS. Eight experienced radiologists interpreted images from 125 selected examinations, 35 with verified findings of cancer and 90 with no finding of cancer. The four display conditions included FFDM alone, 11 low-dose projections, reconstructed digital breast tomosynthesis images, and a combined display mode of FFDM and digital breast tomosynthesis images. Observers rated examinations using the screening BI-RADS rating scale and the free-response receiver operating characteristic paradigm. Observer performance levels were measured as the proportion of examinations prompting recall of patients for further diagnostic evaluation. The results were presented in terms of true-positive fraction and false-positive fraction. Performance levels were compared among the acquisitions and reading modes. Time to view and interpret an examination also was evaluated.

RESULTS. Use of the combination of digital breast tomosynthesis and FFDM was associated with 30% reduction in recall rate for cancer-free examinations that would have led to recall if FFDM had been used alone ($p < 0.0001$ for the participating radiologists, $p = 0.047$ in the context of a generalized population of radiologists). Use of digital breast tomosynthesis alone also tended to reduce recall rates, an average of 10%, although the observed decrease was not statistically significant ($p = 0.09$ for the participating radiologists). There was no convincing evidence that use of digital breast tomosynthesis alone or in combination with FFDM results in a substantial improvement in sensitivity.

CONCLUSION. Use of digital breast tomosynthesis for breast imaging may result in a substantial decrease in recall rate.

[1] Gur, et al. Digital Breast Tomosynthesis: Observer Performance Study. AJR. 2009(193):586-591. http://www.ajronline.org/cgi/content/abstract/193/2/586
Note: The authors acknowledge support from Hologic for funding and cases.

ii. Analysis

The results of the Pittsburgh Reader Study provide additional insight into the interaction of the reader and the tomosynthesis device. Consistent with the Hologic studies, the Pittsburgh study showed that ROC AUC increased with use of 3D tomosynthesis (Figure 8). The Pittsburgh study also showed a lower decrease in the recall rate for non-cancers (Table 14), but an *increase* in the recall rate for cancer patients.

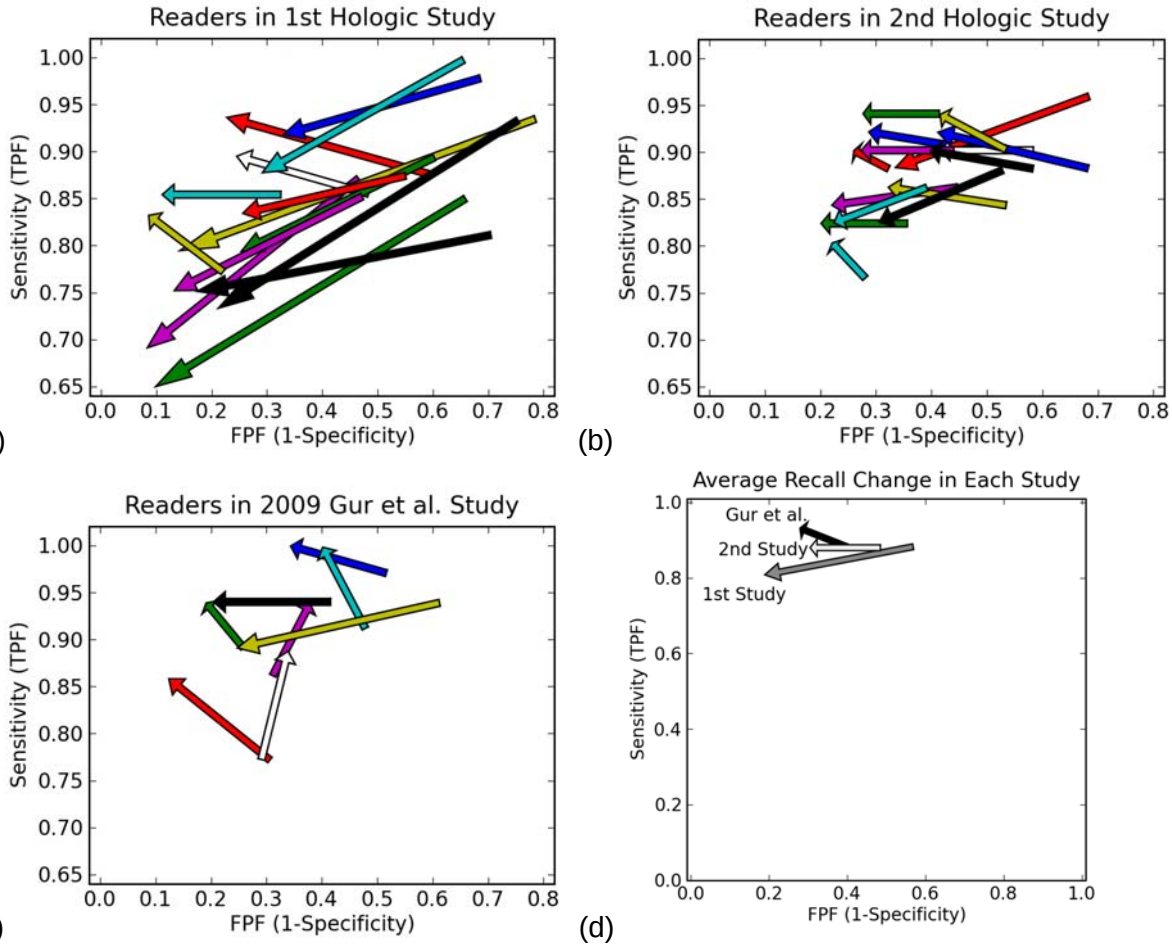


Figure 8. Plots showing how radiologists changed their "recall" operating point (one arrow per radiologist): (a) Reader Study 1, (b) Reader Study 2, and (c) Pittsburgh Study. (d) Illustration of average change in conventional ROC space (i.e., 0.0 to 1.0 on the x- and y-axis). The tail of each arrow indicates radiologist "recall" sensitivity and specificity when using 2D FFDM alone; the head of the arrow indicates a specificity and sensitivity when using 2D with tomosynthesis.

Table 14. Changes in ROC AUC and recall measures with the additional 3D tomosynthesis images (2D plus 3D verses 2D alone for the Hologic Studies). 95% confidence intervals calculated by FDA.

Study	ROC AUC	Cancer Recall	Non-cancer Recall
Reader Study 1	+7% (4%, 11%)	-8% (-15%, 0%)	-38% (-30%, -46%)
Reader Study 2	+7% (4%, 10%)	0%	-19%
University of Pittsburgh Study[2]	+7%, +12%**	+5%	-12%

** The Pittsburgh study used two methods to calculate the change in ROC AUC; both methods were different than the method used by Hologic. The Pittsburgh study used 2-view FFDM plus 2-view DBT (MLO and CC).

[2] Gur et al. "Is an ROC-type Response Truly Always Better Than a Binary Response in Observer Performance Studies?", Academic Radiology 2010 (17)5:639-645

D. Discussion

FDA would appreciate discussion regarding the following topics. Theses issues are considered most relevant to the approval of the device; there may also be areas of differing opinion among members of the review team.

1. Reader Training: The training of the radiologists for Reader Study 2 was influenced by the types of reading errors made during the previous study. In Reader Study 1, the readers were trained not to dismiss lobulated masses even if they were circumscribed; however, based upon review of the dismissed cancer cases approximately half of the readers did not adhere to that training. In Reader Study 2, the readers were again trained not to dismiss lobulated circumscribed masses and their training was reinforced in written format and with further examples. The sponsor noted that based on the scoring of the cancer cases in the dataset, nearly all of the readers in Reader Study 2 had adhered much more closely to the training.

> **FDA Comments**: The enhanced training in Reader Study 2 raises concerns. The reuse of the cancer cases may bias the results in favor of 2D plus 3D, since the training could be specific to cancers that were missed in the initial study. The second study on the same cases would inflate the overall Type I error rate, and the results may not be generalizable to the beyond the set of cancer patients.
>
> However, Reader Study 2 included 15 new readers and a new random sampling of non-cancer cases. The improvement in the ROC curves in both studies indicates that the addition of 3D tomosynthesis increased the discriminative effect for this case set. The shift in operating point matches ROC theory that the discriminative value of a device may be fixed, but through training or experience, readers can adjust their operating points.
>
> **Panel Discussion:** Does the added training and reuse of cancer cases invalidate the Reader Study 2 results?

2. Sequential Reader Design: In both Reader Study 1 and Reader Study 2, data was collected during a sequential read, in which the reader first reviewed and scored the 2D views and then the reader was then shown the additional 3D views and scored the case again.

> **FDA Comments**: In principle, limitations of the sequential reader design could include that the effect of 3D is confounded with extra reading time and that the knowledge that 3D will follow 2D could influence the 2D reading (i.e., cause under- or over-calling of 2D relative to usual practice).
>
> If the comparison was of one imaging mode (e.g., 2D alone) versus another (e.g., 3D alone) non-sequential reading with a memory-wash period would be appropriate. However, the tomosynthesis images provide the ability to further evaluate suspicious areas and search for new findings, which may not necessitate the need for a memory-wash period. The sequential reading design has been accepted in computer aided detection (CAD) studies, where the CAD is intended to provide additional information to the reader. Unlike CAD studies, the “extra reading time” is not spent on the same

images, but on a new set of 3D images. The design appears acceptable for this study given that the 3D views are intended to be used in addition to the 2D views.

The Pittsburgh study used a memory-wash period between the reading modes (in a latin square design with reading session and reader pair). The study also showed an added benefit with tomosynthesis alone, suggesting the sequential reading design did not significantly bias the results. The study also showed that reader performance with FFDM was about the same with or without DBT, suggesting that FFDM readings before viewing DBT were not influenced by knowing that DBT would follow.

Panel Discussion: Does the panel believe the sequential reading scheme significantly affected the study results?

3. Study Design: Additional aspects of the study design that might affect the generalizability of results to screening and diagnostic populations include: retrospective reading, enrichment with recalls and cancer cases, scoring probability of malignancy and BIRADs without the use of priors or clinical history, method of ground truth determination, and patient-based versus lesion-based scoring.

FDA Comments: Many of the issues with the retrospective reader design are considered unavoidable. The study is retrospective and the readers know their actions will not impact the patient. Radiologists also do not have access to prior images or patient history, which could impact their scoring decision. While prospective studies would be ideal, they may be overly burdensome for each sponsor to undertake given the low prevalence of cancer. FFDM devices were approved using retrospective multi-reader multi-case study design.

The enrichment with recalls and cancer cases identified based on standard imaging, including FFDM, would bias the study against the new 3D tomosynthesis device. Statistically, there may be ways to account for this bias, but some of the new techniques may still need to be validated on a clinical dataset. Another possibility would be to perform a study enriched with cancers, but adjust the analysis for the enrichment fraction; to date, FDA has not yet seen sponsors attempt this method.

Many of the confirmed cancer cases had a conventional 2D FFDM examination that resulted in a recall recommendation prior to the 3D tomosynthesis procedure. This presents a bias in favor of the 2D FFDM mode because the selection cancer cases may have excluded cases that might have been recalled with the use of 3D tomosynthesis.

In addition, cases considered negative for cancer were not followed for confirmation. Therefore, some cancer cases may have been misclassified as negative according to the operational definition of ground truth used in the study. However, the likelihood of this may have been low given the low prevalence of cancer. Use an estimate of 0.005 as the cancer rate among screening patients, then less than 1 of the 75 negative screening cases used for the reader study would be expected to have cancer (75*0.005 = 0.375).

In Reader Study 1, the sponsor used patient-based scoring. Reader Study 2 using both patient-based and lesion-based scoring; the similar results suggest that use of lesion-based scoring in Reader Study 1 would not have had much influence on the study.

Panel Discussion: Please discuss if any aspects of the study design are a significant concern for the evaluation of the new 3D tomosynthesis device.

4. Recall Rate: Both Reader Study 1 and Reader Study 2 meet the primary objective for recall rate by showing a significant reduction in the recall rate among screening cases. The recall rate was also stratified by cancer versus non-cancer cases. Both Reader Study 1 and Reader Study 2 showed that 2D plus 3D would reduce the recall rate of non-cancers patients, but *only* Reader Study 2 showed this reduction could be achieved without the loss of cancers.

FDA Comments: The decrease in cancer recall rate in Reader Study 1 raises concerns, but the results for Reader Study 2 suggest that with better training the recall rate could decrease without the loss of cancers.

In Reader Study 2, the cancer recall rate for 2D plus 3D MLO was lower than 2D alone and 2D plus 3D; this result highlights a potential issue with trying to estimate the operating point of readers using a new device and an enriched dataset. Logically, one would expect the results for the 2D plus 3D MLO recall rate to fall between the 2D alone and 2D plus 3D (as with the ROC analysis).

Panel Discussion: (a) Considering only Reader Study 1, would the reduction in the recall rate of cancer patients be acceptable for the sake of significant decreases in the recall rate of non-cancers? (b) Does the data from Reader Study 2 support that 2D plus 3D can reduce the screening recall rate without the loss of cancer patients? (c) Should the sponsor be required to provide additional training, similar to that given in Reader Study 2, to all facilities purchasing the 3D unit?

5. Screening: Both Reader Study 1 and Reader Study 2 showed that using 2D plus 3D would significantly increase in the ROC area under the curve (AUC) and decrease the recall rate of screening patients, but *only* Reader Study 2 showed that these objectives could be achieved without the loss of cancers.

FDA Comments: A risk verses benefit analysis for screening and diagnostic populations can be constructed from the results of the study and assumed prevalences of cancer for these populations. The sponsor conducted their own risk / benefit analysis, which was based on a number of assumptions from the study, cancer prevalence, and radiation risk. However, the study enrichment makes it difficult to generalize the estimates of recall, sensitivity, specificity and any risk / benefit analysis based on these measures. For instance, Reader Study 1 had very high recall rate for 2D of 51.5%, while in practice, the recall rate for 2D may actually be as low as 10%. In addition, 48 of 312 (15%) of cases in the reader study were cancer. The actual prevalence of cancer in screening population may be closer to 5 in 1000.

ROC analysis has previously been accepted for reader studies (e.g., FFDM and CAD) used to evaluate the effectiveness of a new imaging device. The ROC curves measure the performance of a radiologist over a range of decision thresholds (i.e., over a range of sensitivities and specificities); thus, it is useful for measuring the performance of a new imaging modality for which doctors have not yet determined their decision threshold. In clinical practice, the decision thresholds at which a radiologist operates may move along

the ROC curve depending upon experience and the desired tradeoff between sensitivity (cancer detection) and specificity (recall rate).

The consistency of ROC curves between Reader Study 1 and Reader Study 2 suggests that 3D tomosynthesis can help radiologists differentiate between cancer and non-cancer patients. The change in reader operating points between the two studies suggests the importance of training in helping readers understand how to optimally use the new technology. The additional training in Reader Study 2 would be more concerning if there was a large shift in the ROC curve. However, the difference in ROC curves reflects the discriminative effect of the device, where the readers operate reflects the reader's choice on how to apply the additional information.

Reader Study 1, Reader Study 2, and the Pittsburgh study demonstrate that sensitivity and specificity measures are variable and potentially unreliable for predicting future clinical performance of a device. The operating point measures demonstrate the tradeoff expected from the ROC analysis and suggest the importance of training to help readers find their optimal operating point on the ROC curve.

Training will play an important role in determining how radiologists use the device in a premarket study. In Reader Study 1, radiologists were reportedly trained to use the DBT system to eliminate false positives, which greatly increased specificity at the expense of the recall rate of cancer patients. In Reader Study 2, readers received additional 3D training based on the types of errors made during Reader Study 1; shift on the ROC curve to still increase specificity without the decrease in the recall rate for cancer patients. In the Pittsburgh study, radiologists were apparently trained to use the DBT system to find additional cancers, with much less emphasis on false positive reduction.

Panel Discussion: (a) Do the results of Reader Study 1 and Reader Study 2 support the clinical use of 2D FFDM plus 3D tomosynthesis?
(b) Do the results of Reader Study 2 support 2D FFDM plus 3D MLO (1-view) tomosynthesis as another clinical option?

6. Diagnosis: In addition to the primary endpoints, the sponsor provided a "Forced BIRADS" task based comparison of using 2D alone with using 2D plus 3D tomosynthesis in order to support the intended use of the device for "screening and diagnosis".

FDA Comments: Evaluating a "Forced BI-RADS" score is not the same as comparing to diagnostic work-up (e.g., supplementary diagnostic views, spot compressed, and magnification views) or other diagnostic modalities (e.g., ultrasound or MRI). However, in clinical practice, radiologists can make a screening diagnosis of BIRADS 4 or 5 off of a screening exam. In addition to the other results, the "Forced BIRADS" analysis can indicate the diagnostic capability of the system. Since the sponsor is not making claims that 3D tomosynthesis could replace supplementary views or other modalities, the study results suggest the "…and diagnosis" may be appropriate.

Panel Discussion: Is there sufficient data to support 3D tomosynthesis for use as another exam option in the diagnostic setting (i.e., the "…and diagnosis" part of the indications for use)?

IV. Panel Questions

A. Questions for Panel Discussion

FDA would appreciate panel discussion the following aspects of the reader studies. For your convenience, this section consolidates the questions presented in the Discussion section.

1. Reader Training: Does the added training and reuse of cancer cases invalidate the Reader Study 2 results?

2. Sequential Reader Design: Does the panel believe the sequential reading scheme significantly affected the study results?

3. Study design / bias: Please discuss if any other aspects of the study design are a significant concern for the evaluation of the new 3D tomosynthesis device.

4. Recall Rate:

(a) Considering only Reader Study 1, would the reduction in the recall rate of cancer patients be acceptable for the sake of significant decreases in the recall rate of non-cancers?
(b) Does the data from Reader Study 2 support that 2D plus 3D can reduce the screening recall rate without the loss of cancer patients?
(c) Should the sponsor be required to provide training, similar to that given in Reader Study 2, to all facilities purchasing the 3D unit?

5. Screening:

(a) Do the results of Reader Study 1 and Reader Study 2 support the clinical use of 2D FFDM plus 3D tomosynthesis?
(b) Do the results of Reader Study 2 support 2D FFDM plus 3D MLO (1-view) tomosynthesis as another clinical option?

6. Diagnosis: Is there sufficient data to support 3D tomosynthesis for use as another exam option in the diagnostic setting (i.e., the "…and diagnosis" part of the indications for use)?

B. Questions for Ballot Vote

Following the panel discussion, CDRH will ask panel members to vote by ballot on the following three questions related to the approvability of the Selenia Dimensions 3D. Voting will be on the entire indications for use. Multi-part indications will not have each part voted on independently. Please answer them based on your expertise, the information you reviewed in preparation for this meeting, and the information presented at the meeting:

Standard Questions Format:

1. Is there a reasonable assurance that the *[device]* is safe for indication(s) *[X (and Y, etc.)]*?

2. Is there a reasonable assurance that the *[device]* is effective for indication(s) *[X (and Y, etc.)]*?

3. Do the benefits of the *[device]* for indication(s) *[X (and Y, etc.)]* outweigh the risks of the device for indication(s) *[X (and Y, etc.)]*?

Panel members will be asked to electronically vote on each question and to explain their answers. If the panel member answered “no” to any question, he or she will be asked whether changes to labeling, restrictions on use, longer term follow-up, or other controls, would change his or her response.

If the evidence provided is insufficient to allow for any of the determinations, the panel member should state this as the reason for answering “no”. A description of any remedial studies or actions should be given.