Comparison of Breast Tomosynthesis to Conventional Digital Mammography; An Enriched Retrospective Reader Study

Protocol Number: 11-02
Revision 02
December 1, 2011

Sponsor: Hologic, Inc.
35 Crosby Drive
Bedford, MA 01730

Arthur Friedman, Senior Vice President RA/QA/CA
Telephone: (781) 999-7368
Facsimile: (866) 652-8674
E-mail: afriedman@hologic.com

Jeanette Schier-Pugsley, Director, RA
Telephone: (781) 999-7406
Facsimile: (866) 652-8674
E-mail: jschierpugsley@hologic.com

Jennifer Bartoshevich, Manager, CR
Telephone: (781) 999-7764
Facsimile: (866) 652-8674
E-mail: jbartoshevich@hologic.com

The information in this protocol is for Investigators, site personnel, institutional review boards, and health authorities. It will not be disclosed to third parties without written consent from Hologic, except to obtain informed consent from persons entering into the study. Once signed, the terms of the protocol are binding for all parties.

Confidential
# Table of Contents

<table>
<thead>
<tr>
<th>SECTION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXECUTIVE SUMMARY</td>
<td>3</td>
</tr>
<tr>
<td>1.0 Introduction</td>
<td>4</td>
</tr>
<tr>
<td>2.0 Reader Study Design</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Case Selection</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Reader Selection</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Reader Training</td>
<td>8</td>
</tr>
<tr>
<td>2.4 Image Review</td>
<td>10</td>
</tr>
<tr>
<td>2.5 Data Analysis</td>
<td>12</td>
</tr>
<tr>
<td>2.5.1 Primary Endpoint</td>
<td>12</td>
</tr>
<tr>
<td>2.5.1.1 Analysis of ROC Performance</td>
<td>12</td>
</tr>
<tr>
<td>2.5.2 Secondary Endpoints</td>
<td>13</td>
</tr>
<tr>
<td>2.5.2.1 Analysis of ROC Performance for Subjects with Dense Breasts</td>
<td></td>
</tr>
<tr>
<td>2.5.2.2 Analysis of Non-Cancer Recall Rate (Specificity)</td>
<td></td>
</tr>
<tr>
<td>2.5.3 Additional Analyses</td>
<td>13</td>
</tr>
<tr>
<td>2.6 Data Management</td>
<td>14</td>
</tr>
<tr>
<td>2.7 Statistical Tests</td>
<td>14</td>
</tr>
<tr>
<td>2.7.1 Power Analysis for ROC endpoint from previous reader study</td>
<td>14</td>
</tr>
<tr>
<td>2.7.1.1 Number of Reader Study Reviewers</td>
<td>15</td>
</tr>
<tr>
<td>2.7.2 Summary of Results from Previous Reader Study</td>
<td>15</td>
</tr>
<tr>
<td>2.7.3 Recall Rate Endpoints</td>
<td>16</td>
</tr>
<tr>
<td>2.7.4 Additional Analyses</td>
<td>16</td>
</tr>
<tr>
<td>2.7.5 Summary</td>
<td>17</td>
</tr>
</tbody>
</table>
Executive Summary:

This protocol is designed to compare the performance of Tomosynthesis (3D) to FFDM (2D FFDM) in an enriched retrospective reader study. A synthesized 2D image, generated from the tomosynthesis image, will be available to the readers to provide an overview of the anatomy similar to a scout view in CT imaging. Tomosynthesis images used in combination with a synthetic 2D image will be referred to as 3DS. Because the radiation dose for tomosynthesis is approximately the same as FFDM and therefore the risk for a tomosynthesis exam is comparable to that of a FFDM exam, this study will be a non-inferiority study. The primary endpoint will be based on a comparison of the ROC area under the curve for 3DS compared to 2D FFDM imaging.

This study will include images from both fatty and dense breasts with approximately 50% of the cases in each density category. The enriched reader study will include approximately 302 cases. A minimum of ten radiologists will read the images after 2 days of training on the interpretation of 3DS images.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthesized 2D</td>
<td>A 2D image created from the tomosynthesis (3D) images. Requires no additional radiation dose</td>
</tr>
<tr>
<td>3D</td>
<td>Tomosynthesis 3D images</td>
</tr>
<tr>
<td>3DS</td>
<td>Tomosynthesis 3D images plus Synthesized 2D</td>
</tr>
<tr>
<td>2D FFDM</td>
<td>Full field digital mammography (FFDM) – conventional mammography</td>
</tr>
</tbody>
</table>

Primary Endpoint:

- The ROC area under the curve performance for 3DS is non-inferior to that of 2D FFDM

Secondary Endpoint:

- ROC area under the curve for subjects with dense breasts using 3DS is non-inferior to that of 2D FFDM
- The non-cancer recall rate (specificity) for 3DS is non-inferior to that of 2D FFDM
Data Analysis comparing 2D FFDM to 3DS that will be reported:

1) Compare the ROC AUC in subjects with fatty breasts
2) Compare the ROC AUC in calcification versus non-calcification cases
3) Compare recall rates for cancer cases
4) Compare the sensitivity and specificity using the forced BIRADS scores.

1.0 Introduction

Mammography has a well-known limitation that results from the two-dimensional (2D) nature of the imaging. Structures in the breast that lie above and below an object of interest can reduce the conspicuity of the object. This is known as structure noise, caused by superimposed tissue. Superimposed tissue on conventional 2D mammograms may lead to several problems including: 1) obscuring a lesion’s margin, 2) combining or summing to look “suspicious” even when no lesion is present and 3) masking or hiding a breast lesion or cancer.

Tomosynthesis is a new mammographic modality that offers promise to help reduce the problems associated with superimposed tissue. The combination of tomosynthesis plus FFDM has been shown to be superior to FFDM alone based on improved area under the ROC curve and reduced recall rates. Tomosynthesis results in a three-dimensional (3D) set of thin x-ray mammography slices. Relative to conventional mammography, the slices provide improved visibility of objects in thin cross sections in the breast, while reducing the contrast and visibility of objects above or below the cross section of interest.

Tomosynthesis acquisition consists of taking multiple low-dose mammograms at various angles near the normal to the detector while holding the breast compressed in a standard geometry. The acquired images are reconstructed using mathematical algorithms not unlike CT reconstructions, to generate a set of thin slices parallel to the breast platform. The reconstructed slices can be viewed individually or in a movie format. The total radiation exposure to the subject from a tomosynthesis acquisition is similar to that of a conventional 2D FFDM mammogram with tomosynthesis dose about 20% higher than conventional 2D FFDM imaging.

This will be an Enriched Reader Study, following recommendations applicable to the study as outlined in FDA’s Guidance Document Premarket Applications for Digital Mammography Systems; Final Guidance for Industry and FDA, February 16, 2001.

2.0 Reader Study Design

Following the transfer of the complete set of newly acquired, de-identified images and the associated hospital reports (i.e., mammography, diagnostic and pathology
reports, as applicable) from all investigational sites to Hologic, a Reader Study will be initiated to review and analyze the selected subset of images. The criteria by which the subset of images will be selected are described in detail in Section 2.1/Case Selection below.

A minimum of 10 board-certified, MQSA-qualified radiologists will review and score each of the selected image sets according to this protocol (Section 2.4 / Image Review). The radiologists will undergo two days of training to become familiar with reading the tomosynthesis 3D images, as well as become familiar with the workstation that will be used to display the conventional 2D FFDM and tomosynthesis 3D study images. Details of the training are described in Section 2.3 / Reader Training below.

The Reader Study will be performed on Hologic SecurView systems capable of displaying both 2D FFDM and 3D image sets. Each Reader will have access to their own workstation to help ensure independent reviews. The workstation will be equipped with 5 mega-pixel monitors calibrated for mammography image review. Workstations will be located in a quiet, isolated area with controlled lighting simulating the radiologist’s normal working environment. The room lighting will be measured at least once daily with a calibrated light meter and recorded. The lighting in the room will not exceed 10 lux.

The Readers’ image review scores will be recorded from which the subsequent data analysis will be performed.

### 2.1 Case Selection

Cases have been accrued from 22 sites in the United States under IRB approval and with informed consent (Hologic Protocol P09-03). All of the images used for both the 2D FFDM and 3D study images have been acquired on the investigational Selenia Dimensions Digital Breast Tomosynthesis System. This is a new set of images that has not previously been used in any Hologic pilot or pivotal Reader Study. One year follow-up of all cases is being recorded and those cases which are determined to be a cancer within 365 days of the study image date will be considered cancer cases. The 2D FFDM and 3D images were acquired under the same compression minimizing the impact of positioning. The cases are from 2 groups of enrolled subjects.

1. **Screening Subjects** – asymptomatic screening subjects. These women have received a 4-view combination (2D FFDM and 3D) exam in addition to a standard screening mammogram.

---

2. An image set includes the 2D (CC and MLO of both breasts) and 3D (CC and MLO of both breasts) images for one subject ID number.

**Biopsy Subjects**- subjects imaged prior to a breast biopsy, BIRADS 3, 4 and 5 subjects. These subjects have received a 4-view combination (**2D FFDM** and **3D**) exam prior to breast biopsy. The breast biopsy subjects were not limited to subjects with a mammographically detected lesion. A subject with a clinical, MRI or ultrasound detected lesion were also eligible.

Fatty and dense breast cases will be used in this Reader Study. The density scoring was performed at the acquisition site by two radiologists reading the standard of care **2D FFDM** images. The standard of care **2D FFDM** images are the subject's clinical mammography images and are not used in the Reader Study. If the site radiologists agreed on the categorization of fatty (BIRADS 1 or 2) or dense BIRADS 3 or 4), their density scores were used to categorize the case as fatty or dense. If the two site radiologists disagreed on the categorization, a third independent radiologist scored the case to determine the final categorization.

During case collection, each subject’s images were read by two radiologists. The first radiologist scored the standard of care **2D FFDM** images with access to priors and patient history. This was the standard clinical interpretation. A second radiologist read and scored the investigational images (**2D FFDM** and **3D**) again with access to prior images and clinical history. If either radiologist requested further imaging (recalled the subject) then the subject was called back and the required imaging was performed.

A subset of cases will be selected to create the Reader Study case Set and includes the following groups:

- **Negative screening cases** – about 126 cases – negative by **2D FFDM** and **2D FFDM plus 3D** reader at acquisition site (BIRADS 1 or 2 score considered negative)
- **Recalled cases** – about 24 cases – recalled by **2D FFDM** or **2D FFDM plus 3D** reader at acquisition site
- **Benign biopsy cases** about 76 cases – pathology proven benign cases (30% of the cases present as calcification cases with no associated mass)
- **Cancer cases** – about 76 cases – pathology proven malignant cases (30% of the cases present as calcification cases with no associated mass)

Each case will be assigned to only one of the 4 groups.

The cancer cases will include both calcification and non-calcification lesions. The cases will be selected from a randomized list of cancer cases; however, cases with a mass of greater than or equal to 2.5 cm in size will be excluded. For the cancer cases, the first 24 (first 12 in dense breasts and first 12 in fatty breasts) cancers presenting as calcifications and the first 52 (first 26 in dense breasts and first 26 in fatty breasts) cancers presenting as non-calcification lesions (masses or distortion) will be included in the Reader Study. This will provide a mix of cancer types similar to that described in the DMIST study in which 30% of the cases were DCIS and 70%
were invasive\textsuperscript{4} assuming that DCIS will be detected primarily as calcifications while invasive cancers will primarily be detected by the presence of a mass or distortion. The size and types of cancer cases are consistent with recommendations of the March 4th, 2008 Advisory Panel’s recommendations\textsuperscript{5}. Cancers with both a mass and calcifications will be included with the non-calcification cancer cases.

The cases in each group will be placed in group-specific chronological order and given a consecutive number from 1 up to N where N is the number of cases in that group. An independent statistician will generate and provide a random order list of these cases to determine the selection order. It is possible that some cancer lesions were found at the acquisition sites upon review of the 3D images. These cancers may be included if they are among the first 24 calcification or 52 non-calcification lesions selected from the randomized list. In some cases several cancer lesions may be found in the same subject, however for the case selection, these cases will only be placed into a Reader Study group once. The cases will be 50% fatty (BIRADS 1 or 2) and 50% dense (BIRADS 3 or 4) to allow comparison of these groups and to provide sufficient cases for the dense breast secondary endpoint analysis.

The non-cancer cases will also be listed in a consecutive manner and selected with the same method. The mixture of recalled, negative screening, cancer and benign biopsy cases is designed to provide a case set enriched in both recalls and cancers to allow evaluation of the study endpoints in a study set of reasonable size. Negative screening and benign biopsy cases are included to balance the recall and cancer cases and provide a mixture of cases. The number of cancer cases has been increased from approximately 50 in the previous reader studies to 76 in this study. The number of cancer cases has been increased primarily to increase the power for the ROC comparison.

Cases acquired from the enrolled Biopsy Subject group will be assigned to either the cancer or benign Reader Study group based on their pathology results.

A screening case recalled by either the 2D FFDM or 2D FFDM plus 3D site reader will be defined as a recall case unless after recall the subject proceeds to a biopsy. In that situation, the case will be defined as either a benign biopsy or cancer Reader Study case, depending on the pathology results. Negative screening cases are screening cases that were not recalled by the 2D FFDM or 2D FFDM plus 3D site reader.


2.2 Reader Selection

A minimum of 10 board certified radiologists with a range of experience based on the volume of cases read per year will be selected to read and score the study images. The goal will be to get roughly equal numbers of radiologists in the three groups defined below:

- High Volume Reader - defined as radiologists who read more than 5000 mammograms per year.
- Medium Volume Reader - defined as radiologists who read more than 3000 up to 5000 mammograms per year.
- Low Volume Reader - defined as radiologists who read 3000 or fewer mammograms per year.

A survey questionnaire will be sent to prospective Readers to determine the fraction of their professional time devoted to breast imaging. Hologic has established that as a requirement for being considered a Reader for the trial, all Readers will have provided Hologic with documentation of experience in reading full-field digital mammograms (FFDMs) acquired on the Hologic Selenia or Hologic Dimensions FFDM systems and reading experience on the Hologic SecurView review workstation. All of the proper documentation for each Reader to participate in the study will be maintained on file in accordance with Hologic record retention policy. Readers will be compensated for the 6 days required to complete the study.

The readers will also include a range of experience with tomosynthesis from those using it clinically to those without experience using tomosynthesis.

Inclusion Criteria for Readers:

1) Must be an MSQA qualified Radiologist

2) Must have at least 1 year experience reading digital mammograms

3) All potential Readers must pass the pass/fail training data set thresholds in order to be a valid Reader in the actual study. Details are described in Section 2.3.

4) Must be familiar with reading Hologic 2D FFDM images on the Hologic SecurView review workstation.

2.3 Reader Training

The goal of the Reader training is to assure that the radiologists participating in the Reader Study are trained in the interpretation of the tomosynthesis images. A radiologist with expertise in breast tomosynthesis will conduct the training of the radiologists.

Prior to the start of training the readers will be given a 2D FFDM assessment set consisting of 60 cases having approximately 14 cancer, 14 benign biopsy, 6 recalled...
and 26 negative screening image sets. Readers will be required to pass performance thresholds described below to be included as a Study Reader.

The radiologists will receive 2 days of training in the evaluation of tomosynthesis images. A synthesized 2D image reconstructed from the tomosynthesis images will be included with each of the 3D images for review. For each view, mediolateral oblique (MLO) and cranio-caudal (CC) the corresponding synthesized 2D image for the same view will be shown with the 3D images (3DS image set)). Similar to the training in previous reader studies it will be emphasized that lesions presenting with lobulated margins on the 3D images should be recalled as many of these lesions have been found to be cancers. The training will also emphasize that the synthetic 2D images alone will not be used for diagnosis, scoring will be based on the appearance of the lesion on the 3D images.

Training will consist of providing the trainees with an initial overview of tomosynthesis images by presenting and discussing a series of approximately 50 cases. These cases will include approximately

- one-third cancer cases,
- one third recalled or superimposed tissue cases, and
- a mixture of benign biopsy and negative screening cases.

The initial review will take approximately 3 hours and will consist of the trainer presenting the selected cases to demonstrate the range of typical anatomy and abnormalities identified with tomosynthesis. Training will also consist of a hands-on session at the workstation to provide the Readers with an overview of its tomosynthesis-specific functionality.

Following the initial training, all Readers will be required to independently review and score two different Case Sets. The Initial Case Set will consist of approximately 40 image sets having approximately 10 cancer, 10 benign biopsy, 15 negative screening and 5 recall image sets. Following the independent read of this Case Set, the trainer will review the image set with the group.

On the second training day, all Readers will be required to independently review and score the Final 3DS Assessment Case Set including 60 cases with the same mix of cases used for the 2D FFDM assessment set. Readers will be required to pass performance thresholds shown below. These 2D FFDM and 3DS final assessment sets will not include the same cases.

For these 2D FFDM and 3DS Final Assessment sets, the Readers will be required to meet the Pass Criteria, which is:

**Recall rate threshold** – this threshold is based on the recall rate for the 26 negative cases included in the 3DS and 2D FFDM and assessment sets.

- The recall rate threshold for readers scoring the 3DS assessment set is the mean recall rate of all readers plus 1.65 times the standard deviation of the
number of cases recalled for each reader. Any reader that recalls more cases than this threshold will be excluded.

- The recall rate threshold for readers scoring the 2D FFDM assessment set is the mean recall rate of all readers plus 1.65 times the standard deviation of the number of cases recalled for each reader. Any reader that recalls more cases than this threshold will be excluded.

**Cancer Detection Threshold** – this threshold is based on the forced BI-RADS scores for the 14 cancer cases included in the 3DS and 2D FFDM and assessment sets. A cancer is considered detected if the case is scored a BI-RADS score of 4a, 4b, 4c or 5.

- The cancer detection threshold for readers scoring the 3DS assessment set is the mean detection rate of all readers minus 1.65 times the standard deviation of the number of cases detected for each reader. Any reader that detects fewer cases than this threshold will be excluded.

- The cancer detection threshold for readers scoring the 2D FFDM assessment set is the mean detection rate of all readers minus 1.65 times the standard deviation of the number of cases detected for each reader. Any reader that detects fewer cases than this threshold will be excluded.

To be included as a Reader in the Reader Study, all Pass Criteria for Final Case Sets must have been achieved, as well as completion of the entire training session. Any Reader failing this Final Case Set review will continue to review and score the Reader Study Case Set, but this Reader’s scores will not be included in the final study data analysis.

### 2.4 Image Review

Following the successful completion of the Reader training, the Readers will independently review and score each of the image sets of the Reader Study Case Set. Each Reader will work at a designated workstation.

The readers will read the 2D FFDM and 3DS images as shown in the flow chart below. During the first read (Session 1) readers will read 1/2 of cases using 2D FFDM and 1/2 of the cases with 3DS. They will return one month later for a second reading (Session 2). The same readers will read all cases using both methods. Prior 2D FFDM images will not be used in this study. A synthesized 2D image created from the 3D data set will be provided to the readers along with the 3D images for both the CC and MLO views when the corresponding 3D images are viewed.
Figure 1
Flow Chart for Enriched Reader Study

Reader Study Case Set – 302 Cases

2D FFDM read

3D S read

One Month Delay

3D S read

2D FFDM read

Data Analysis

Note: there is a score recorded for each mode shown in the above diagram, totaling 2 scores for each case.

Scoring will be lesion based. The radiologists will mark the position of any lesions suspicious enough to be recalled using an ellipse annotation on the images. For 3DS lesion scoring, the lesions will be marked on one slice of the tomosynthesis image near the center of the lesion. For 3DS scoring all lesion scoring and marks will be on the 3D images. Lesions will not be marked or scored on the synthetic 2D images. Electronic scoring of the lesion X, Y and slice number (if applicable) will also be recorded by a scribe. After marking the position the radiologist will be asked to identify the lesion type (calcification, distortion, mass or a combination of these lesion types) and then provide a forced BIRADS score (1 to 5) and a probability of malignancy (POM) score (0 to 100%). Each lesion will have a POM and BIRADS score. These scores will be treated as independent scores. It is also possible that a reader will recall a lesion but feel the lesion is likely benign or superimposed tissue and still give the lesion a low forced BIRADS score and low POM score. If the radiologist does not mark any lesion on a case it is assumed the radiologists would not recall the case and they will be asked to give a BIRADS 1 or 2 score for the case and a POM.

In the analysis of lesion detection, a lesion will be considered detected if the center X and Y coordinates of the lesion marked by the readers are within 2 cm of the center coordinates marked by an expert reader prior to the study. For images
marked on 3D images the slice number must be within +/- 10 slices (+/- 1 cm) from the slice marked as the lesion center by the expert radiologist. If there are questions regarding the correct lesion location the ellipse marked by the reader may be used to resolve these questions. For example if there is a multi-focal lesion or a calcification cluster spread over a large area of the breast one radiologist may mark the whole area as one cancer while another may mark several lesions. The ellipse marked on the image will help to resolve these lesion identification questions. Training will encourage the radiologists to use separate marks if there is normal breast tissue between lesions. A reader is required to correctly mark the lesion on only 1 view, so for example a lesion correctly marked on the RCC view but not marked on the RMLO view would still be considered a correct lesion detection.

For the ROC analysis the lesion with the highest POM score will be used as the POM score for a given case. For each cancer lesion an expert radiologist will mark the position of the lesion on both the 2D FFDM and 3D images prior to the start of the Reader Study. The expert will use all available imaging reports and case report forms from the acquisition site to determine the lesion position and size.

2.5 Data Analysis

2.5.1 Primary Endpoint

2.5.1.1 Analysis of ROC Performance

The radiologist’s POM score for each case will be used to generate the 2D FFDM and 3D ROC curves.

**Endpoint Analysis** – A multi-reader, multi-case ROC analysis will be used to compare 3D to 2D FFDM. The areas under the curve (AUC) will be used to compare ROC performance. 3D will be considered non-inferior to 2D FFDM if the lower limit one-sided 95% CI for the difference in AUCs (3D minus 2D FFDM) is greater than -0.05. That is, our null hypothesis is that the AUC for 3D is 0.05 less than the AUC for 2D. A difference of 0.05 is considered a clinically significant difference.

ROC performance has been selected for the primary endpoint because it is the best method of testing a new imaging modality. In the early stages of a new technology radiologists have not yet selected the proper thresholds for recalling a case. Once a technology is mature and radiologists have sufficient experience it would make sense to look at the binary choice of recall versus no recall; however, in the early stages of a technology using ROC methodology is superior since it tests a new device over a range of recall thresholds.

---

2.5.2 Secondary Endpoint

2.5.2.1 Analysis of ROC Performance in Subjects with Dense Breasts

The radiologist’s POM case based score will be used to generate the 2D FFDM and 3DS ROC curves for women with dense breasts (defined as BIRADS density score of 3 or 4).

**Endpoint Analysis** – A multi-reader, multi-case ROC analysis will be used to compare 3DS to 2D FFDM. The areas under the curve (AUC) will be used to compare ROC performance. 3DS will be considered non-inferior to 2D FFDM if the lower limit one-sided 95% CI for the difference in AUCs (3DS minus 2D FFDM) is greater than -0.05. A difference of 0.05 is considered a clinically significant difference.

2.5.2.2 Analysis of Non-Cancer Recall Rate (Specificity): 3D$_S$ compared to 2D FFDM

A case based comparison of non-cancer recall rates will be made by comparing the total number of cases recalled with 3D$_S$ to that for 2D FFDM. This analysis will use the initial BIRADS score of 0 (recall) and 1 or 2 (no recall).

**Endpoint Analysis** – The analysis of recall rate will be for all non-cancer cases including negative screening, recalled and benign biopsy cases. A bootstrapping method with replacement will be used to compare average recall rates among all Readers’ pooled results for 3D$_S$ to that for 2D FFDM. If the upper limit to the one-sided 95% Confidence Interval for the difference (3D$_S$ minus 2D FFDM) in recall rates among non-cancers is less than 0.05 (delta of 5%), 3D$_S$ will be considered non-inferior to that of 2D FFDM.

2.5.3 Additional Analyses

1. Report ROC performance and non-cancer recall rate for calcification cases versus non-calcification cases.
2. Report the Sensitivity and Specificity based on forced BIRADS scores. A score of 4a, 4b, 4c, or 5 will be considered a positive score.
3. Compare ROC performance for subjects with fatty breasts (BIRADS density score of 1 or 2).
4. Compare recall rates for cancer cases for 2D FFDM and 3D$_S$. Note this analysis will be case based but a correct cancer recall will require correct lesion location and type. The fraction of cancer lesions correctly marked will also be reported (no p-value for this comparison will be reported due to within patient correlations).
2.6 Data Management

The data management company (DYAD Systems, Cambridge Massachusetts) will provide reports for the final training session and the reading session in order to reduce missing data or missing case report forms for each Reader.

2.7 Statistical Tests

Primary Endpoint - ROC Analysis:

Shown below is the statistical power analysis from a previous reader study. Based on this analysis and this study’s results, 50 cancer cases, 260 negative cases and 10 readers was demonstrated to be sufficient to test the ROC endpoint. For the proposed Reader Study, Hologic is increasing the number of cancer cases to 76 and leaving the number of non-cancer cases approximately the same with 226 cases total. Thus with the increased cancer cases the proposed Reader Study will have much greater than 80% power to test the non-inferior ROC endpoint.

A description of a power analysis for a previous study is shown below. The results of a pilot study are also shown. The previous study could be used to develop a power analysis but there were important differences between that study and the current study including: 1) synthetic 2D algorithms have been improved significantly and the previous study used an early version of this software, 2) the current study also has a secondary endpoint for a dense breast ROC analysis and there is insufficient data from the previous study to develop a power analysis for this subset of subjects, 3) the previous pilot did not use lesion marking. For these reason we report the previous power analysis based on published estimates of reader performance but have increased the number of cancers to allow for the dense breast analysis. The sample size was determined based upon the assumption that the AUC of the two methods were both equal and in the range of 0.80 to 0.85.\textsuperscript{7} This is based upon the assumption that the AUC with the synthetic 2D algorithm and lesion based scoring will be similar to the assumed AUC for the patient level scoring of the previous study.

2.7.1 Power Analysis for ROC Endpoint from Previous Reader Study

For this analysis, a total of 50 to 60 cancers were expected to be included in the reader study. Cancer detection and lesion classification were compared using Receiver Operator Characteristics (ROC) curve analyses. The ROC performance (Area under the ROC curve (AUC)) was compared for the review of the 3D\textsubscript{s} images and the review of the 2D FFDM images. The number of reviewers required for the Reader Study was determined as follows:

2.7.1.1 Number of Reader Study Reviewers from Previous Reader Study

The sample size (number of readers) for the Reader Study was calculated by the method of Obuchowski\(^8\). It was based on the following assumptions:

- AUC for control modality: between 0.78 and 0.89 (Selenia PMA P080003)
- Delta AUC: 0.05
- Number of positive subjects: 50
- Number of negative subjects: 260
- Variance between readers, \(\sigma^2_b\): 0.0009 (Beam, Layde and Sullivan)
- Correlation between areas when a set of readers evaluate the same subject sample using different diagnostic tests, \(r_b\): 0.82 (Obuchowski, 1995)
- Within subject variance, \(\sigma^2_w\): 0.0001 (Obuchowski, 1995)
- Correlations: \(r_1 = 0.4356\), \(r_2 = 0.3272\), \(r_3 = 0.2917\) (Obuchowski, 1995)
- Power to detect a difference in AUC of 80%.

The method of Hanley and McNeil (1982) was used to estimate the variance component attributable to subject sample sizes, 0.0006.

Based upon these values, 10 readers would suffice for any combination of control AUC and numbers of positive and negative subjects in the ranges above.

2.7.2 Summary of Results from Previous Reader Study

Hologic has performed a reader study with 13 radiologists using a similar case distribution of approximately 50 cancers, 48 benign and 210 screening cases. This study included all types of breasts (dense and fatty) and all lesion types (non-calcification and calcifications). A synthesized 2D image was shown along with the 3D images. The results are presented to further support the sample size calculations. Figure 2 shows the overall ROC results for all cases. This study was a comparison of 2D FFDM to 3DS. The ROC AUC was 0.849 for 2D FFDM and 0.875 for 3DS, p-value 0.19. The difference in ROC AUC was not significantly different for 3DS compared to 2D FFDM in this study.

\(^8\) Obuchowski NA. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. Acad Radiol 1995; 2(suppl 1):S22–S29.
For the primary ROC endpoint analysis, the proposed Reader Study will include 76 cancers compared to 50 in the previous study and this will provide sufficient power to see small changes between 2D FFDM and 3D and comparison of the two methods for subjects with dense breasts. In the previous study a difference of 0.04 in ROC AUC would have resulted in a significant difference in ROC AUC.

2.7.3 Recall Rate Endpoint

For the non-cancer recall rate endpoint an analysis of the previous reader study recall rates was done using a bootstrapping method with replacement and treating the readers as random readers. Based on the results of that analysis the proposed design with 250 non-cancer cases will have power to detect an average difference of 4.6% in non-cancer recall rates.

2.7.4 Additional Analyses

The cancer recall rate will be defined as the fraction of cancer cases recalled with 3D and 2D FFDM. Any cancer lesion marked and given a forced BIRADS and probability of malignancy (POM) score will be considered to have been recalled if the
cancer lesion and lesion type was correctly marked. In addition to the case based analysis, the lesion level cancer recall rate will be reported.

An analysis of the previous reader study’s cancer recall rates was done using a bootstrapping method with replacement and treating the readers as random readers. Using this method it was estimated that using 50 cancer cases the standard error for the pooled estimate was 3.3%. If this analysis is extended to 76 cancer cases, the predicted standard error would be 2.7%.

2.7.5 Summary

Based on the sample size calculations presented above, approximately 76 pathology confirmed cancer cases, 76 pathology confirmed benign cases and approximately 150 negative screening and recalled cases combined will be sufficient to demonstrate the study’s primary endpoint and secondary endpoints.