

BIostatISTICS Consulting, LLC

10606 WHEATLEY STREET / KENSINGTON, MD 20895

CONFIDENTIAL

Statistical Analysis Plan

for

FUJIFILM Medical Systems U.S.A., Inc.

protocol

**FMSU2017-002B: A Multi-Reader Multi-Case Controlled Clinical Trial to Evaluate the
Comparative Accuracy of Fujifilm DBT plus S-View versus FFDM Alone in the Detection
of Breast Cancer – A Pivotal Study**

5 April 2018

Statistical Analysis Plan

for

FUJIFILM Medical Systems U.S.A., Inc.

protocol

FMSU2017-002B: A Multi-Reader Multi-Case Controlled Clinical Trial to Evaluate the Comparative Accuracy of Fujifilm DBT plus S-View versus FFDM Alone in the Detection of Breast Cancer – A Pivotal Study

Approved by:

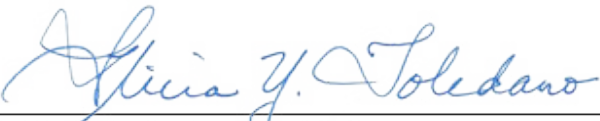

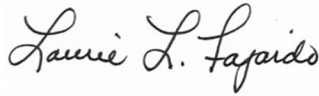
	Date: 2018-04-12
Alicia Y. Toledano, Sc.D. Study Statistician Biostatistics Consulting, LLC	
	Date: 4/11/2018
Nicole Calabrese, M.S. Contract Project Manager FUJIFILM Medical Systems U.S.A., Inc.	
	Date: 4/21/2018
Laurie Fajardo, M.D., M.B.A, F.A.C.R, F.S.B.I. Principal Investigator Radiology Consultant	

TABLE OF CONTENTS

1. Introduction	5
1.1. Study Endpoints	5
1.1.1. Primary Endpoint	5
1.1.2. Secondary Endpoints	6
2. Study Design.....	7
2.1. Study Population (Cases)	7
2.2. Study Radiologists (Readers).....	8
3. Test Methods	9
3.1. Reference Standard	9
3.2. Randomization	9
3.3. Image Review Procedures.....	10
3.4. Image Interpretation Results	10
3.5. Lesion Matching (Scoring).....	11
3.6. Blinding / Masking.....	11
4. Statistical Methods	12
4.1. Study Samples (Analysis Sets).....	13
4.2. Treatment Assignment	13
4.3. Multiple Centers (Pooling).....	13
4.4. Derived Variables.....	13
4.5. Subgroups.....	16
4.6. Analysis of Study Endpoints and Important Subgroups	17
4.6.1. Primary Endpoint	17
4.6.2. Secondary Endpoints	18
4.6.3. Multiple Comparisons.....	20
4.7. Test Reproducibility.....	25
4.8. Interim Analyses	25
4.9. Safety Monitoring Analyses (Adverse Events).....	25
4.10. Sample Size Calculations	25
4.11. Data Quality Review	29
5. Results to be Reported.....	30
6. Regulatory and Administrative Information.....	32
7. References	33
Appendix 1.....	34

LIST OF TABLES

Table 1. Per-Subject POM and BI-RADS Scores Requiring Correct Lesion Localization.....	15
Table 2. Per-Subject Recall Scores Requiring Correct Lesion Localization	16
Table 3. Hypotheses That May Be Tested	22
Table 4. Estimated Power for Primary Endpoint: Non-inferior AUC	28

LIST OF FIGURES

Figure 1. Graphical Approach to Protect the Study's Type 1 Error Rate from Inflation	24
--	----

ABBREVIATIONS

ANOVA	analysis of variance
AUC	area under the receiver operating characteristic (ROC) curve
BCL	Biostatistics Consulting, LLC
BI-RADS	Breast Imaging Reporting and Data System <i>BI-RADS® is a registered trademark of the American College of Radiology</i>
CC	craniocaudal
CI	confidence interval
CRF	case report form
DBT	digital breast tomosynthesis
eCRF	electronic case report form
FDA	U.S. Food and Drug Administration
FFDM	full field digital mammography
Fujifilm	FUJIFILM Medical Systems U.S.A., Inc.
MLO	mediolateral oblique
MQSA	Mammography Quality Standards Act
MRMC	multi-reader multi-case
POM	probability of malignancy
ROC	receiver operating characteristic
SAP	statistical analysis plan
SD	standard deviation
S-View	synthesized view

1. Introduction

This document provides the statistical analysis plan (SAP) for FUJIFILM Medical Systems U.S.A., Inc. (Fujifilm) protocol FMSU2017-002B, which is a retrospective, multi-reader, multi-case (MRMC) pivotal study to be conducted with an enriched sample of approximately 300 cases and approximately 18 board-certified and Mammography Quality Standards Act (MQSA)-qualified radiologists with a range of experience who will be trained to read and evaluate Fujifilm digital breast tomosynthesis (DBT) and synthesized view (S-View) images. Each radiologist will review full field digital mammography (FFDM) and DBT plus S-View images for each case in a counterbalanced design with an approximately four (4) week memory washout period. The purpose of the pivotal reader study is to evaluate the comparative accuracy of Fujifilm DBT plus S-View versus FFDM in the detection of breast cancer. In particular, this study will evaluate whether the DBT plus S-View read is non-inferior to the FFDM alone read. The results of pivotal study FMSU2017-002B are intended to support a regulatory submission for the Fujifilm ASPIRE Cristalle DBT plus S-View system.

This SAP is based on approved protocol FMSU2017-002B Final Version 1.0 dated 14 March 2018. If the protocol is amended in a manner that requires this SAP to be revised, Fujifilm and Biostatistics Consulting, LLC (BCL) will finalize the revised SAP before locking the database for the primary analysis. If there is a conflict between the protocol and this SAP, the language in this SAP as approved by BCL, Fujifilm, and the study Principal Investigator shall prevail.

1.1. Study Endpoints

1.1.1. Primary Endpoint

The primary endpoint is non-inferior per-subject average area under the receiver operating characteristic (ROC) curve (AUC) for DBT plus S-View versus FFDM, based on probability of malignancy (POM) scores requiring correct lesion localization.

The study will be considered to have successfully demonstrated safety and effectiveness of the Fujifilm ASPIRE Cristalle DBT plus S-View system if the per-subject average AUC for DBT plus S-View is statistically significantly non-inferior to the average AUC for FFDM at the $\alpha = 0.05$ significance level, for non-inferiority margin $\delta = 0.05$. This will be established if

the lower limit of the two-sided 95% CI for the difference in average AUC for DBT plus S-View minus FFDM lies entirely above -0.05 .

1.1.2. Secondary Endpoints

The secondary endpoints are:

1. Non-inferior and/or superior (lower) per-subject average recall rate for all non-cancer cases for DBT plus S-View versus FFDM, based on recall scores, using non-inferiority margin $\delta = 0.05$.
2. Non-inferior per-subject average recall rate for DBT plus S-View versus FFDM for all cancer cases, based on recall score, using non-inferiority margin $\delta = 0.10$.
3. Non-inferior and/or superior per-subject average sensitivity for DBT plus S-View versus FFDM, based on BI-RADS scores requiring correct lesion localization, using non-inferiority margin $\delta = 0.10$.
4. Superior per-subject average AUC for DBT plus S-View versus FFDM, based on POM scores and requiring correct lesion localization.
5. Non-inferior per-subject average specificity for DBT plus S-View versus FFDM, based on BI-RADS scores.
6. Superior (lower) per-subject average recall rate for DBT plus S-View versus FFDM for all follow-up proven non-cancer recall cases.
7. Non-inferior per-lesion average sensitivity for masses, masses with calcifications, focal asymmetries, and/or architectural distortions for DBT plus S-View versus FFDM, based on BI-RADS scores requiring correct lesion localization, using non-inferiority margin $\delta = 0.10$.
8. Non-inferior and/or superior per-lesion average sensitivity for calcifications for DBT plus S-View versus FFDM, based on BI-RADS scores requiring correct lesion localization, using non-inferiority margin $\delta = 0.10$.
9. Non-inferior (margin $\delta = 0.05$ for AUC, 0.10 for other performance metrics) and/or superior average AUC and/or other performance metric(s) for DBT plus S-View versus FFDM for subjects with dense breasts (BI-RADS breast composition categories c. The breasts are heterogeneously dense, which may obscure small masses and d. The breasts are extremely dense, which lowers the sensitivity of mammography).

Estimates and corresponding 95% confidence intervals illustrating precision in the estimates will be provided for, at a minimum: per-subject average recall rate for all non-cancer cases and for all cancer cases, per-subject average sensitivity, and per-subject average specificity. (Note: This information is provided for AUC as part of evaluation of the study's primary endpoint.). Sensitivity, specificity, and recall rate each provide additional information needed to understand the expected impact of using DBT plus S-View on clinical practice. Their analysis and reporting, including confidence intervals, is thus required to provide a complete description of device performance.

Multiple comparisons. To protect the study's Type 1 error rate from inflation we will use the iterative graphical approach described in Bretz, et al.¹ to sequentially reject hypotheses.

2. Study Design

Protocol FMSU2017-002B is a retrospective, MRMC study of Fujifilm DBT plus S-View to be conducted with an enriched sample of 300 cases obtained from multiple image acquisition centers on Fujifilm protocol FMSU2013-004A "Acquisition of Digital Mammography and Breast Tomosynthesis Images for Clinical Evaluation of Fujifilm Digital Breast Tomosynthesis," and approximately 18 radiologist readers with varying experience levels some of whom have limited experience reading 2D synthetic images. The study employs a fully factorial, counterbalanced crossover design in which all readers review images from all cases in two (2) visits separated by a memory washout period of approximately four (4) weeks. Each reader will read half the cases as FFDM and the other half as DBT plus S-View during Visit 1, and the complementary FFDM and DBT plus S-View images during Visit 2.

2.1. Study Population (Cases)

Protocol FMSU2017-002B will include an enriched sample of 300 cases obtained from multiple image acquisition centers on protocol FMSU2013-004A. As part of the consent process for protocol FMSU2013-004A subjects agreed that image data and supporting documentation could be used for future research and investigations. All cases for this pivotal MRMC reader study will meet the following eligibility inclusion and exclusion criteria:

Inclusion Criteria

- Eligible subjects under protocol FMSU2013-004A, defined as female subjects with known true clinical status and with complete FFDM and DBT examinations, in which

there is sufficient anatomical coverage, sufficient contrast, and no significant motion or other artifacts, as determined by the image-acquisition sites.

- Meet none of the exclusion criteria under protocol FMSU2013-004A.

Exclusion Criteria

- Subjects who are in violation of protocol FMSU2013-004A.
- Subjects who meet exclusion criteria under Fujifilm protocol FMSU2013-004A.
- Subjects with unknown clinical status.
- Any subject whose positive mammogram was not read during the truthing process will not be considered for the pivotal reader study.

Case selection. The study sample was selected by BCL per the approved “Case Selection Specifications for FUJIFILM Medical Systems U.S.A., Inc. protocol FMSU2017-002B: A Multi-Reader Multi-Case Controlled Clinical Trial to Evaluate the Comparative Accuracy of Fujifilm DBT plus S-View versus FFDM Alone in the Detection of Breast Cancer – A Pivotal Study (Revision 1)”, dated and approved 7 March 2018. Cases were selected from subjects with images acquired in the N-Mode dose setting, as specified in the Indications for Use for the ASPIRE Cristalle Digital Breast Tomosynthesis Option (P160031). Cases were selected based on levels of the following factors: 1) Reference standard status, 2) Presence of calcifications (yes or no), 3) Breast composition (density), and to the extent possible given the pool of available cases, 4) Acquisition site. Backup cases to be used in the event that selected cases were not available (for example, image files become corrupted) also were identified, matched as closely as possible to the values of the selection factors for each selected case. Images from all cases used on the pivotal study will have passed quality control review by the Fujifilm team.

Demographic and clinical characteristics. Demographic and clinical characteristics were obtained on Fujifilm protocol FMSU2013-004A (see **Appendix 1** for relevant form pages). The sample includes 60 cancer cases and 240 non-cancer cases comprised of 48 benign cases, 72 recall cases, and 120 normal cases. Cases without biopsy are from the screening or recall enrollment pathways and have one-year negative imaging follow-up.

2.2. Study Radiologists (Readers)

Approximately 18 radiologists will participate as study readers. Readers may be radiologists of varying experience levels, from both community and academic practices, some of whom have

limited experience reading 2D synthetic images. Reader information is recorded on a dedicated two-part (2-part) questionnaire (**Appendix 1**).

Qualifications. All readers must be board-certified and Mammography Quality Standards Act (MQSA)-qualified for both FFDM and DBT interpretation.

Training. Readers will receive approximately three (3) hours of training in the evaluation of DBT plus S-View images. Training will also consist of a hands-on session at the workstation to provide the readers with an overview of its DBT-specific functionality. For each view, mediolateral oblique (MLO) and craniocaudal (CC) 2D FFDM and the corresponding S-View image for the same view will be shown with the DBT images. The training will also emphasize that the S-View images alone will not be used for diagnosis, and scoring will be based on the appearance of the lesion on the DBT images.

3. Test Methods

3.1. Reference Standard

The reference standard for cancer and benign cases is biopsy proof. The reference standard for recall and normal cases is one-year follow-up imaging (320 to 455 days inclusive). The truthers' lesion type(s) and location(s) for all cancer cases in both modalities (FFDM and DBT plus S-View) will be recorded on an electronic case report form (eCRF; **Appendix 1**).

3.2. Randomization

Randomization was performed by BCL per the approved "Randomization Specifications for FUJIFILM Medical Systems U.S.A., Inc. protocol FMSU2017-002B: A Multi-Reader Multi-Case Controlled Clinical Trial to Evaluate the Comparative Accuracy of Fujifilm DBT plus S-View versus FFDM Alone in the Detection of Breast Cancer – A Pivotal Study," dated and approved 09 March 2018.

The 300 cases were randomly allocated into four (4) sets of 75 cases each, case subsets A, B, C, and D, each with 15 cancer cases, 12 benign cases, 18 recall cases, and 30 normal cases. Allocation was balanced to the extent possible on presence of calcifications, breast composition (fatty or dense), and image acquisition site using optimal nonbipartite random matching.² Each Session will comprise three (3) days of reading; we refer to the first part of the Session as Part 1, and the remainder as Part 2.

We used the principle of Latin squares to reduce bias associated with reading order.³ This was achieved by counterbalancing case sets and reading condition (FFDM, DBT plus S-View) in the four portions of each Session (Part 1 Group 1, Part 1 Group 2, Part 2 Group 1, Part 2 Group 2) to ensure that:

- Interpretation within each Part of the Session is half FFDM, half DBT plus S-View (that is, we place this constraint on the study design);
- Each Session uses all four (4) case sets A, B, C, D;
- Each case set is interpreted once as FFDM and once as DBT plus S-View, on separate Sessions;
- Each of FFDM, DBT plus S-View is interpreted at a different Part and Group combination between Sessions; and
- Each case set is interpreted at a different Part and Group combination between Sessions.

Reading order was randomly determined for each reader. The Wald-Wolfowitz runs test⁴, at statistical significance level $\alpha = 0.05$, was used to ensure that for each reader the null hypothesis that the distribution of cancer cases in the list is random was not rejected in favor of the two-sided alternative hypothesis that this distribution is not random.

3.3. Image Review Procedures

Study readings are scheduled to occur at International HealthCare, LLC (Norwalk, CT) between 6 April 2018 and 24 May 2018. Each reader will read both FFDM and DBT plus S-View images for each case, separated by a memory washout period, on the ASPIRE Bellus II workstation.

3.4. Image Interpretation Results

Readers will be prompted by scribes, who will enter each reader's responses in the reader eCRF (**Appendix 1**). For each case on each read, the reader will first note whether there are mammographic findings. If the answer to this question is "no" the reader will be asked to provide a BI-RADS assessment category of 1 or 2, a probability of malignancy (POM) score in 0% through 100%, and a recall decision of "no." If the reader answers "yes" to whether there are mammographic findings the reader will be asked to confirm an initial BI-RADS assessment

category of 0, and will then provide detailed information for up to three (3) suspicious findings (reader lesions):

- Location (including breast [right or left], diagram location [1 – 9 or combinations when the finding is in multiple diagram sections] within view [right CC, left CC, right MLO, left MLO], and coordinates for each of CC and MLO [N/A if not seen on that view; or X, Y, and for DBT only, Slice])
- Type, as mass, asymmetry, calcification, architectural distortion, or other with description. The reader may check all that apply.
- Forced BI-RADS assessment category 1, 2, 3, 4, or 5
- POM in 0% through 100%

The reader will then be asked for her or his overall recall decision (yes or no), forced BI-RADS assessment category, and POM score, for the case.

In cases with mammographic findings, consistency of BI-RADS scores, POM scores, and recall decisions will not be forced – for example, readers will be permitted to use the full range of POM scores for a finding no matter what BI-RADS score they assign to it.

3.5. Lesion Matching (Scoring)

An expert not associated with diagnosing cases at the image acquisition sites or serving as a study reader will perform lesion matching to determine whether the location and type of any reader findings match a lesion annotated by the truther. Lesion matching will be performed for all malignant lesions in cancer cases. The lesion matcher's results will be recorded on an eCRF (**Appendix 1**).

3.6. Blinding / Masking

The readers will be told that the samples of cases do not represent a standard screening population, and will be blinded to the actual distribution and nature of the set of images they will be asked to review. Readers will be masked to the reference standard and image acquisition interpretations (under Fujifilm protocol FMSU2013-004A) for each case. Readers will not have access to prior mammograms or other clinical information. All readers will perform their interpretations independently.

4. Statistical Methods

Informed consent. By approving this SAP, Fujifilm confirms the following: All subjects whose images were acquired under FMSU2013-004A and selected for this study were consented. As part of the consent process, subjects agreed that image data and supporting documentation could be used for future research and investigations. Each reader will be consented before initiating the reader study.

Masking to protect identities. Study case identification numbers and study reader numbers will be assigned to all cases and readers, and used to protect their identities in statistical analysis and in reporting of results.

Statistician not blinded. Because the reader data on lesion locations only includes Slice for DBT, the statistician will not be blinded to reading condition.

General conventions: Descriptive summaries. Baseline descriptive summaries will include the distribution of demographic characteristics and clinical characteristics, including characteristics specific to malignant and, if appropriate, benign lesions. We also will provide summaries across readers of the per-subject number of findings, BI-RADS scores, POM scores, and per-subject recall scores, for the FFDM and DBT plus S-View readings. These may be cross-classified by, for example, presence of malignant lesions. Categorical variables (such as cancer type and breast tissue composition) generally will be summarized using frequencies and proportions or percentages, while continuous variables generally will be summarized using means and standard deviations (SDs), and/or medians and quartiles or ranges. Missing values generally will be reported as such in these descriptive summaries.

General conventions: Statistical inferences. Uncertainty in estimates of diagnostic accuracy will be quantified through confidence intervals (CIs). Unless otherwise noted, statistical inference procedures (hypothesis tests, CIs) are two-sided with significance level $\alpha = 0.05$ and corresponding confidence level 0.95. Statistical inferences for proportions (for example, sensitivity and specificity) may use the binomial distribution or other exact methods rather than normal approximations, for example, when sample sizes are small and/or when proportions are close to zero or one. Results will be presented by reader using reader numbers to mask reader identities, and averaged across readers.

4.1. Study Samples (Analysis Sets)

Intent-to-diagnose population. The intent-to-diagnose population comprises each reader's interpretation of each study case in each modality (FFDM, DBT plus S-View).⁵ Missing interpretations are not anticipated; in the event that any interpretations are missing in the study database they will be accounted for when reporting study results. We plan to include all readers' interpretations of all cases in both modalities in the analysis set.

If any protocol deviations or violations are reported to the statistician, the statistician will evaluate them to determine their impact on the integrity of the study database, and will determine whether any affected data points should be excluded from primary analysis (that is, whether primary analysis should occur in a modified intent-to-diagnose population).

Unit of analysis. The primary unit of analysis on this study is the subject (case). Malignant lesion is the secondary unit of analysis for per-lesion sensitivity.

4.2. Treatment Assignment

This is a retrospective study for which imaging and clinical management occurred prior to case selection. All cases will be evaluated the same way by all study readers, such that there are no treatment assignments or treatment groups.

4.3. Multiple Centers (Pooling)

Fujifilm obtained images from multiple centers. The protocol for data submission, quality review passed by all images, and reference standard status determination for all images used in the pivotal MRMC study were common. Cases will be pooled across enrolling centers for interpretation on the pivotal MRMC study using common interpretation protocol and eCRFs, and results of interpretation sessions will be monitored. The scoring of (lesion matching for) reader interpretations will follow a common process. Results for any particular reader therefore will be pooled across enrolling centers.

4.4. Derived Variables

Per-subject BI-RADS, POM, and recall scores requiring correct lesion localization will be derived as shown below. The general principle is that even at the subject level, credit is only given for identifying a subject with cancer if the reader marks findings in at least one location

with cancer. Findings that do not match the location of a malignant lesion are ignored for cancer cases in the per-subject analyses but may be reported, generally in an appendix.

Per-Subject scoring: POM and BI-RADS. The primary endpoint is per-subject AUC based on POM scores requiring correct lesion localization. Secondary endpoints include per-subject sensitivity requiring correct lesion localization and specificity based on BI-RADS categories. When computing sensitivity and specificity based on BI-RADS, a score of 4 or 5 constitutes a positive test result. A cutoff score of BI-RADS 3 or higher may also be used to compute the sensitivity and specificity in secondary analyses. Scores for use in these analyses will be derived by the statistician as summarized in **Table 1** on page 15.

Per-Subject scoring: Recall. Secondary endpoints include per-subject recall rate for non-cancer cases and separately for cancer cases requiring correct lesion localization, based on a separate yes/no question. Scores for use in this per-subject analysis will be derived by the statistician as summarized in **Table 2**.

True Positive, False Negative, True Negative, and False Positive. In per-subject analysis of sensitivity and specificity:

- A true positive (TP) occurs when a case contains one or more cancerous lesions and the per-subject BI-RADS score requiring correct lesion localization is 4 or 5.
- A false negative (FN) occurs when a case contains one or more cancerous lesions and the per-subject BI-RADS score requiring correct lesion localization is 1, 2, or 3.
- A true negative (TN) occurs when a case does not have any cancerous lesions and the per-subject BI-RADS score is 1, 2, or 3.
- A false positive (FP) occurs when a case does not have any cancerous lesions and the per-subject BI-RADS score is 4 or 5.

A cutoff score of BI-RADS 3 or higher may also be used to compute the sensitivity and specificity in secondary analyses. When computing recall rates requiring correct lesion localization, a *recall* occurs when a case has per-subject recall score equal to yes.

Table 1. Per-Subject POM and BI-RADS Scores Requiring Correct Lesion Localization

Reference standard	Reader's interpretation	Per-Subject POM and BI-RADS
No malignancies in this case	No findings in this case	POM: Same as POM recorded by the reader for the case. BI-RADS: Same as BI-RADS category recorded by the reader for the case. <i>From Initial Mammographic Findings form page.</i>
	One or more findings in this case	POM: Overall POM recorded by the reader for the case. BI-RADS: Overall BI-RADS category recorded by the reader for the case. <i>From Overall Patient Recall form page.</i>
One or more malignancies in this case ^{1,2}	No findings in this case	POM: Same as POM recorded by the reader for the case. BI-RADS: Same as BI-RADS category recorded by the reader for the case. <i>From Initial Mammographic Findings form page.</i>
	Findings in this case, but no findings matching the location(s) of any proven malignancies in this case	POM: Assigned as the higher of 0 or, for readers who do not assign POM 0 to any case in a reading modality, the minimum POM score assigned by that reader in that modality. BI-RADS: Assigned as category one (1).
	One or more findings correctly matching the location(s) of any proven malignancies in this case	POM: Highest POM score recorded by the reader for any of these matched findings. BI-RADS: Highest BI-RADS category recorded by the reader for any of these matched findings.

¹If the case contains more than one malignant lesion, the reader will get credit for identifying the case as having one or more proven malignancies even if the reader does not identify all of the proven malignancies in the case. For example in a bilateral case, the reader would get credit for identifying the case even if the reader marks findings in only one breast.

²The POM scores and BI-RADS categories for any reader findings in this case that do not match the location(s) of any proven malignancies will be ignored in the per-subject analysis, which requires a single POM score and single BI-RADS category per subject conditional on whether the subject does or does not have proven malignancies.

Table 2. Per-Subject Recall Scores Requiring Correct Lesion Localization

Reference standard	Reader's interpretation	Per-Subject Recall Score
No malignancies in this case	No recall (initially or overall)	Same as recall recorded by the reader for the case, that is, no recall.
	Recall (initially and overall)	Same as recall recorded by the reader for the case, that is, recall.
One or more malignancies in this case ¹	No recall (initially or overall)	Same as recall recorded by the reader for the case, that is, no recall.
	Recall (initially and overall) and Findings in this case, but no findings matching the location(s) of any proven malignancies in this case	Assigned as no recall.
	Recall (initially and overall) and One or more findings correctly matching the location(s) of any proven malignancies in this case	Same as recall recorded by the reader for the case, that is, recall.

¹If the case contains more than one malignant lesion, the reader will get credit for recalling the case even if the reader does not identify all of the proven malignancies in the case. For example in a bilateral case, the reader would get credit for recalling the case even if the reader marks findings in only one breast as long as the overall decision is to recall the subject.

4.5. Subgroups

Analyses of per-subject recall rate for non-cancer cases and specificity are limited to the subgroup of cases without cancer. Analyses of per-subject sensitivity and per-subject recall rate for cancer cases are limited to the subgroup of subjects with cancer. We may also 1) analyze recall rate in the subgroups of non-cancer recall, normal, and benign cases; 2) perform per-lesion analysis of sensitivity in subgroups defined by lesion type (masses with or without calcifications, focal asymmetries, and/or architectural distortions in one subgroup, and calcifications in another subgroup); and/or 3) analyze AUC, sensitivity, specificity, and/or recall rate in the subgroups of women with dense or non-dense breasts.

4.6. Analysis of Study Endpoints

4.6.1. Primary Endpoint

The primary endpoint on this study is non-inferior per-subject average AUC requiring correct lesion localization. Primary analysis will not involve pooling across study radiologists, to allow for heterogeneity across them. We will estimate AUCs for each reader in each review condition (FFDM, DBT plus S-View) based on per-subject POM scores requiring correct lesion localization derived as in **Section 4.4**, above. The non-inferiority margin for this endpoint is $\delta = 0.05$.

We will provide graphs of each reader's ROC curve for each review condition. For each reader, the non-parametric (trapezoidal) AUC for the FFDM read, the DBT plus S-View read, and the difference between them, will be presented. Statistical inferences will account for correlations arising from having all study readers interpret all study cases. We plan to compare AUCs between reading conditions using the standard MRMC analysis of variance (ANOVA) method of Obuchowski and Rockette⁶ adjusted for estimation in the F^* test statistic⁷, to ensure generalization of the study results to both the population of readers and the population of cases. Two-sided 95% CIs will be used to quantify uncertainty in the within-modality estimates and the between-modalities differences.

Modeling framework. Let A_{ij} be an estimate of the AUC in modality i ($i = 1$ for FFDM, 2 for DBT plus S-View) for the j^{th} radiologist ($j = \text{reader } 1, \dots, R$ for $R = 18$). We consider the effects of radiologists to be random, because interest extends beyond the radiologists on this study to a larger population of potential radiologists from which these radiologists are a sample. Obuchowski and Rockette⁶ model these estimates using mixed effects ANOVA, as

$$A_{ij} = \mu + \alpha_i + b_j + (ab)_{ij} + e_{ij}$$

where

- μ is the overall AUC across the populations of readers and cases,
- α_i is the fixed effect of modality,
- b_j is a random effect for reader with expectation 0 and variance σ_b^2 with random effects for different readers independent of each other,

- $(\alpha b)_{ij}$ is a random effect for the interaction of modality and reader with expectation 0 and variance σ_{ab}^2 also with random effects for different readers independent of each other, and
- e_{ij} is random error with expectation 0, variance $\sigma_c^2 + \sigma_w^2$ for σ_c^2 the case sample variance and σ_w^2 the within-reader variance, and covariance
 - $r_1 \sigma_c^2$ for two AUCs from the same reader in different modalities,
 - $r_2 \sigma_c^2$ for two AUCs from different readers in the same modality, and
 - $r_3 \sigma_c^2$ for two AUCs from different readers in different modalities.
- The random effects b_j , $(\alpha b)_{ij}$, and e_{ij} are independent of each other.
- When the readers review the case sample only once in each modality $(\alpha b)_{ij}$ and e_{ij} are not identifiable, and we cannot separate σ_{ab}^2 from $\sigma_c^2 + \sigma_w^2$.

We will obtain the average AUC within each modality and its standard error, and the average difference in AUC for DBT plus S-View – FFDM and its standard error. We will use these to compute corresponding two-sided 95% CIs for the average AUC within each modality and for their difference, all referencing a Student's t -distribution with degrees of freedom adjusted for estimation in Obuchowski and Rockette's⁶ adjusted F-statistic.⁷

The study will be considered to have successfully demonstrated safety and effectiveness of the Fujifilm ASPIRE Cristalle DBT plus S-View system if the per-subject average AUC for DBT plus S-View is statistically significantly non-inferior to the average AUC for FFDM at the $\alpha = 0.05$ significance level, for non-inferiority margin $\delta = 0.05$. This will be established if the lower limit of the two-sided 95% CI for the difference in average AUC for DBT plus S-View minus FFDM lies entirely above the negative of the non-inferiority margin, -0.05 .

4.6.2. Secondary Endpoints

When analyzing secondary endpoints per-subject POM, BI-RADS, and recall scores requiring correct lesion localization will be derived as described in **Section 4.4**. Analyses of secondary endpoints also will use standard MRMC ANOVA methods^{6, 7} to compare performance metrics for DBT plus S-View versus FFDM, and two-sided 95% CIs to quantify uncertainty.

The secondary endpoints are:

1. Non-inferior and/or superior (lower) per-subject average recall rate for all non-cancer cases, based on recall scores, using non-inferiority margin $\delta = 0.05$.
2. Non-inferior per-subject average recall rate for all cancer cases, based on recall score, using non-inferiority margin $\delta = 0.10$.
3. Non-inferior and/or superior per-subject average sensitivity, based on BI-RADS scores requiring correct lesion localization, using non-inferiority margin $\delta = 0.10$.
4. Superior per-subject average AUC, based on POM scores and requiring correct lesion localization.
5. Non-inferior per-subject average specificity, based on BI-RADS scores.
6. Superior (lower) per-subject average recall rate for all follow-up proven non-cancer recall cases.
7. Non-inferior per-lesion average sensitivity for masses, masses with calcifications, focal asymmetries, and/or architectural distortions, based on BI-RADS scores requiring correct lesion localization, using non-inferiority margin $\delta = 0.10$.
8. Non-inferior and/or superior per-lesion average sensitivity for calcifications, based on BI-RADS scores requiring correct lesion localization, using non-inferiority margin $\delta = 0.10$.
9. Non-inferior (margin $\delta = 0.05$ for AUC, 0.10 for other performance metrics) and/or superior average AUC and/or other performance metric(s) for subjects with dense breasts (BI-RADS breast composition categories c. The breasts are heterogeneously dense, which may obscure small masses and d. The breasts are extremely dense, which lowers the sensitivity of mammography).

Estimates and corresponding 95% confidence intervals illustrating precision in the estimates will be provided for, at a minimum: 1) per-subject average recall rate for all non-cancer cases and 2) for all cancer cases, 3) per-subject average sensitivity, and 4) per-subject average specificity. Analysis of each of these will be performed using standard MRMCA ANOVA methods.^{6,7} Sensitivity, specificity, and recall rate each provide additional information needed to understand the expected impact of using DBT plus S-View on clinical practice. Their analysis

and reporting, including confidence intervals, is thus required to provide a complete description of device performance.

Analysis evaluating 5) superiority of per-subject average AUC will be performed similarly to analysis of the Primary Aim. Superiority will be established if the lower limit of the two-sided 95% CI for the difference in average AUC for DBT plus S-View – FFDM lies entirely above zero (0).

Analysis of 6) per-subject recall rate for all recall cases will be performed similarly to analysis of per-subject recall rates for all non-cancer cases. If analysis is performed in the subgroup of recall cases, complementary analysis will be performed in the subgroup of normal cases and in the subgroup of benign cases.

We plan to use methods for clustered data from MRMC studies that take into account the correlation between lesions in the same case when analyzing lesion-level sensitivity in 7) the subgroup of soft tissue lesions (masses, masses with calcifications, focal asymmetries, and/or architectural distortions) and 8) the subgroup of calcifications. In particular, Rao and Scott's⁸ method for estimating proportions from clustered data will be used to obtain estimates for each reader in each reading condition, and Obuchowski's⁹ extension of this to a pair of correlated proportions will be used to estimate the variance-covariance matrix of all possible pairs of proportions. The usual Obuchowski and Rockette^{6,7} MRMC method will then be applied to perform inferences that generalize to the population of readers and the population of cases while also taking into account within-case correlations between lesions. Analysis will be performed in either both subgroups, or in neither subgroup.

Analysis 9) in the subgroup of subjects with dense breasts will be performed similarly to corresponding analysis above. If analysis is performed in the subgroup of women with dense breasts, complementary analysis will be performed in the subgroup of women with non-dense breasts.

4.6.3. Multiple Comparisons

We will use a graphical approach to illustrate relationships among endpoints and protect the study's type 1 error rate from inflation.^{1,10} The testing strategy is shown in a figure with vertices (nodes) for each hypothesis to be tested and directed paths (arrows) between vertices. All study hypotheses to be formally tested as potential marketing claims are shown in the graph.

Hypotheses are denoted $H_1, H_2, \dots, H_k, \dots, H_m$, where m is the total number of hypotheses to be tested. Each hypothesis to be tested is allocated initial endpoint specific alpha, α_k , and the sum $\alpha_1 + \alpha_2 + \dots + \alpha_k + \dots + \alpha_m = \alpha$ for the study overall, 0.05. Initial endpoint-specific alpha can be 0 for endpoints of lesser importance; hypotheses for these endpoints will receive alpha from hypotheses that are rejected and direct alpha toward them. Each directed path is assigned a weight between 0 and 1 indicating how much of the endpoint-specific alpha moves along the path when a hypothesis is rejected. The sum of weights leaving any hypothesis equal to 1, that is, all of the preserved alpha is used in receiving hypotheses. Paths may include a loop-back feature whereby if only one of a pair of looping hypotheses is rejected at its endpoint-specific α_k , that α_k loops back to the other hypothesis to increase its endpoint-specific alpha. Testing in a strategy with loop-back can start with any vertex that has initial endpoint-specific $\alpha_k > 0$, and all such vertices can be tested until one is found for which the null hypothesis is rejected; then testing follows the arrows. Finally, conditional passing of alpha is shown by paths with negligible weights epsilon (ϵ); this places higher priority on other hypotheses until those have been tested.

The graph is continually updated each time a null hypothesis is successfully rejected, as follows:

1. Pass α_k from successful H_k according to the path weights.
2. Eliminate the vertex for H_k .
3. Connect all incoming arrows to outgoing arrow tails of the deleted vertex.
4. Adjust path weights based on relative weights of previous parts of path. Maintain:
 - a) Sum of endpoint-specific $\alpha_k = \alpha$, and
 - b) Sum of outgoing weights from each vertex = 1.
5. If a new path duplicates an existing path, combine them and add their weights.

The hypotheses that Fujifilm is interested in testing on this study are shown in **Table 3** on the following page.

Table 3. Hypotheses That May Be Tested

Hypothesis	Endpoint	Null Hypothesis	Alternative Hypothesis	Non-inferiority margin, delta
H ₁	AUC	Inferior	Non-inferior	0.05
H ₂	Per-subject recall rate for all non-cancer cases	Inferior	Non-inferior	0.05
H ₃	Per-subject recall rate for all cancer cases	Inferior	Non-inferior	0.10
H ₄	Per-subject sensitivity	Inferior	Non-inferior	0.10
H ₅	Per-subject recall rate for all non-cancer cases	Equal	Superior	N/A
H ₆	Per-subject specificity	Inferior	Non-inferior	0.05
H ₇	AUC	Equal	Superior	N/A
H ₈	Per-subject sensitivity	Equal	Superior	N/A
N/A = Not applicable.				

General considerations for employing the graphical approach to protect the study's type 1 error rate from inflation, as presented in protocol section 12.6, are:

- H₁ corresponds to the study's primary endpoint. Testing will use the full study alpha: $\alpha_1 = \alpha = 0.05$. Hypothesis testing for secondary endpoints will only proceed if H₁ is rejected, in which situation α_1 will be passed to one or more of H₂, H₃, and/or H₄ through path weights w_{12} , w_{13} , and w_{14} , respectively. These weights will sum to 1, and one or two of them may be 0.
- H₂, H₃, and H₄ correspond to higher priority secondary endpoints. Their initial endpoint-specific alphas are zero, because they are only tested if the study's primary aim is met. If that occurs their endpoint-specific alphas are updated to $\alpha_k = w_{1k} \times \alpha_1$. Each of H₂, H₃, and H₄ may pass alpha to either of the others, and this may involve loop-back. One or more of H₂, H₃, and/or H₄ also may pass alpha to one or more of H₅ through H₈, and this may be conditional on first testing all of H₂, H₃, and H₄.
- H₅ through H₈ correspond to secondary endpoints with lower priority and/or likelihood of success. Their initial endpoint-specific alphas are zero. Each of these may receive alpha

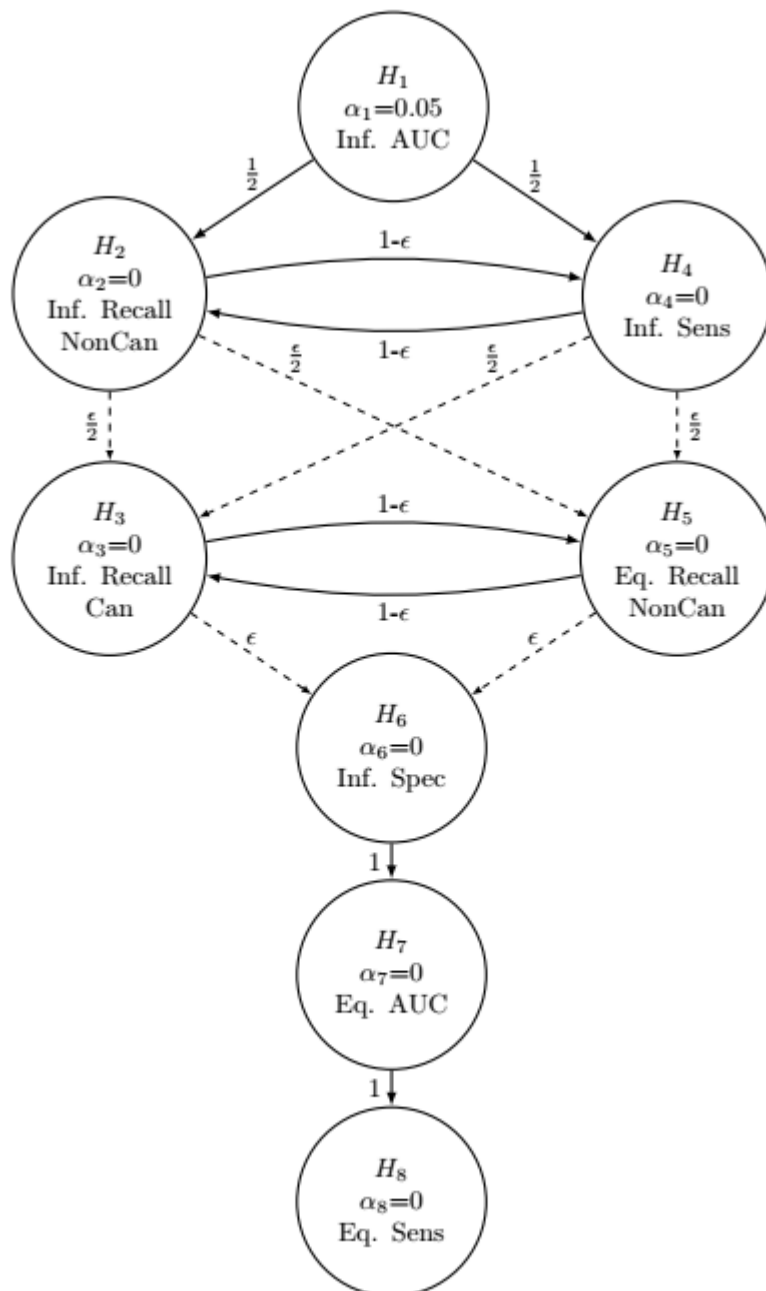
from H_2 , H_3 , and/or H_4 . Each of H_5 through H_8 may pass alpha to the others, and this may involve loop-back.

- Paths with weight 0 will not appear in the final graphic.

By approving this SAP, Fujifilm and the study's Principal Investigator confirm that this pre-specified statistical analysis plan (SAP) provides final path weights determined through discussion between Fujifilm, the study's Principal Investigator, and BCL. Path weights have been finalized prior to receiving any study reading data, as follows:

- H_1 passes alpha equally to each of H_2 and H_4 .
- H_2 and H_4 pass alpha to each other with path weights $1 - \epsilon$ (epsilon) and to each of H_3 and H_5 with path weights $\epsilon/2$.
 - The pair of infinitesimal path weights $\epsilon/2$ ensure that both H_2 and H_4 are tested before testing H_3 and H_5 .
 - The loop-back provision between H_2 and H_4 ensures that if either the p-value for testing H_2 : inferior non-cancer recall rate or the p-value for testing H_2 : inferior sensitivity is less than $\alpha/2$ (0.025), the other endpoint can be tested at $\alpha = 0.05$ without inflating the study's type 1 error.
- H_3 and H_5 pass alpha to each other with path weights $1 - \epsilon$ (epsilon), and to H_6 with path weight ϵ .
- H_6 passes its entire alpha to H_7 (path weight 1), and
- H_7 passes its entire alpha to H_8 (path weight 1).

The graphical approach with path weights is provided in **Figure 1** on the following page.



Inf. = Inferior. AUC = Area under the receiver operating characteristic (ROC) curve.
 NonCan = Non-Cancer. Sens = Sensitivity. Can = Cancer. Eq. = Equal. Spec = Specificity.

Figure 1. Graphical Approach to Protect the Study's Type 1 Error Rate from Inflation

4.7. Test Reproducibility

Test reproducibility will not be evaluated on this pivotal MRMC study.

4.8. Interim Analyses

No interim analyses of study endpoints are planned.

4.9. Safety Monitoring Analyses (Adverse Events)

No adverse events are anticipated on this pivotal MRMC study using retrospective cases for which medical management has already been planned and carried out. Readers also are unlikely to report any adverse events. Any adverse events that are reported to BCL will be described.

4.10. Sample Size Calculations

A sample size of at least 60 cancer cases, 240 non-cancer cases, and 18 readers was selected for this study. Sample sizes were calculated to show non-inferior AUC on a per-subject basis (primary endpoint), using results from a recently conducted pilot study. Inflation of type 1 error associated with using these results to size the pivotal study is likely to be negligible because: 1) the readers used in the pivotal and pilot studies are mutually exclusive, 2) a random sample of the pilot cases are used in the pivotal study (that is, there is no selection bias), 3) the size of the pivotal study is large relative to the size of the pilot study, and 4) training of readers will not be modified according to the pilot study results.

We used the method of Obuchowski^{11,12} to determine the number of readers required in a fully crossed design to provide 80% power at statistical significance level $\alpha = 0.05$ for the multi-reader, multi-case (MRMC) F*-test when the number of positive cases (cancers) is 60 and the number of non-cancer cases is 240. Calculations were made based on the following assumptions:

- *Endpoint:* Difference in average area under ROC curve (A) for DBT plus S-View versus FFDM
- *Significance level:* $\alpha = 0.05$
- *Target power:* 80%
- *Non-inferiority margin:* $\delta = 0.05$

- *Null hypothesis:* The average area under the ROC curve for DBT plus S-View, A_2 , is inferior to the average area under the ROC curve for FFDM, A_1 , by an amount equal to the non-inferiority margin:

$$H_0: A_2 \leq A_1 - \delta.$$

- *Alternative hypothesis:* The average area under the ROC curve for DBT plus S-View, A_2 , is *not* inferior to the average area under the ROC curve for FFDM, A_1 :

$$H_A: A_2 > A_1 - \delta.$$

- Average area under ROC curve for FFDM, A_1 , = 0.80
- For calculating power,
 - Average area under the ROC curve for DBT plus S-View, A_2 , = $A_1 - \delta = 0.75$ under the null hypothesis of inferiority H_0 , and
 - $A_2 = A_1 = 0.80$ under the alternative hypothesis of non-inferiority, H_A .
- *Statistical procedure:* Two-sided 95% confidence interval for $A_2 - A_1$ constructed as

$$\hat{A}_2 - \hat{A}_1 \pm (t_{\text{crit}} \times \text{SD}).$$

The “hat” (^) indicates an estimated quantity.

- t_{crit} is the 0.975th quantile of Student's t -distribution with $R - 1$ degrees of freedom, for R the number of study readers. The Hillis⁷ adjustment to degrees of freedom is not made in these power calculations because using $R - 1$ is more conservative. For convenience of notation, degrees of freedom are not explicitly denoted in t_{crit}
- SD is the standard deviation of $\hat{A}_2 - \hat{A}_1$
- For sample size calculations we use the duality between a two-sided 95% confidence interval and a one-sided hypothesis test for $A_2 - A_1 > \delta$ at significance level $\alpha = 0.025$. The test statistic is:

$$t^* = \frac{(\hat{A}_2 - \hat{A}_1) - (-\delta)}{\sqrt{\frac{1}{R(R-1)} \sum_{j=1}^R \left\{ (\hat{A}_{ij} - \hat{A}_{ir})^2 - (\hat{A}_i - \hat{A}_r)^2 \right\} + 2\hat{\sigma}_c^2(\hat{r}_2 - \hat{r}_3)}}$$

for reading conditions $i = 1, 2$ and readers $j = 1, \dots, R$, r_2 the correlation between two AUCs from different readers in the same modality (both DBT plus S-View or both FFDM), r_3 the correlation between two AUCs from different readers in different modalities, and σ_c^2 the variance because cases are a sample from a larger population.

- *Confidence level:* 95%, that is, two-sided significance level is $\alpha = 0.05$.
- *Criterion:* The null hypothesis will be rejected in favor of the alternative if $t^* > t_{\text{crit}}$.

Power depends on the number of cancer cases, number of non-cancer (benign, recalled, and normal) cases, and number of readers, R through the variance of the difference⁶,

$$\text{var}(\hat{A}_2 - \hat{A}_1) = \frac{2}{R} \{ \sigma_{ab}^2 + \sigma_w^2 + \sigma_c^2 [1 - r_1 + (R - 1)(r_2 - r_3)] \},$$

which includes variance components σ_{ab}^2 for the interaction between reader and reading condition, σ_w^2 for within-reader variance (that is, when the same reader interprets the same case sample in the same reading condition), σ_c^2 because the cases are a sample, correlation of A within reader, between reading conditions, r_1 , and the difference in correlation of A between reader, within reading condition versus between conditions, $r_2 - r_3$. Numbers of cancer and non-cancer cases enter this variance through σ_c^2 from a binormal approximation¹². The closed-form expression for σ_c^2 facilitates use when the ratio of non-cancer cases to cancer cases in the study being planned may differ from that ratio in the pilot study from which estimates of other parameters are obtained. We used data from the pilot study to estimate the variance components and correlations in $\text{var}(\hat{A}_2 - \hat{A}_1)$. To be conservative the component σ_{ab}^2 was estimated as $\sigma_b^2 \times (1 - r_b)$, because the unbiased estimate from ANOVA may be negative. Values used in power calculations were:

- $\sigma_{ab}^2 = \sigma_b^2 \times (1 - r_b) = 0.0004$, from $\sigma_b^2 = 0.0016$ and $r_b = 0.73$.
- $\sigma_w^2 = 0.0001$.
- σ_c^2 , calculated using a binormal approximation.¹² This variance depends on the values of A_1 and A_2 , and on the numbers of cancer and non-cancer cases in the sample.
- $r_1 = 0.53$.
- $r_2 - r_3 = 0.02$.

Estimates of power for MRMC studies are highly dependent on the assumptions above. We therefore also obtained the number of readers required to provide 80% protected power allowing for possible attrition or parameter misspecification:

- 15% attrition rate for either cases or readers.
- Decrease in the comparator metric, A_1 , to 90% of its assumed value.

- If power decreased when a parameter value increased, we increased said parameter by 1/2 on the measurement scale (50%; by 125% on the variance scale correspondingly).
- If power decreased when a parameter value decreased, we decreased said parameter by 1/3 on the measurement scale.

Table 4 on the following page shows that a study with 18 readers, 60 cancers, and 240 non-cancers, provides at least 80% protected power.

Table 4. Estimated Power for Primary Endpoint: Non-inferior AUC

	Power (%)*
Initial values**: $A_1 = 0.80$, $\delta = 0.05$, n = 60 cancers + 240 non-cancers, R = 18 readers, $\sigma_{ab}^2 = 0.0004$ from $\sigma_b^2 = 0.0016$ and $r_b = 0.73$, $r_1 = 0.53$, $(r_2 - r_3) = 0.02$ and $\sigma_w^2 = 0.0001$	92
0.90 A_1	91
0.85n	89
0.85R	88
2.25 σ_{ab}^2	85
0.67 r_1	88
2.25 σ_w^2	91
1.5($r_2 - r_3$)	89
<p>*Power calculated using Student's t distribution with R – 1 degrees of freedom and rounded down to nearest whole percent.</p> <p>**A_1 = Area under the ROC curve with FFDM. δ = Non-inferiority margin. σ_{ab}^2 = variance for interaction between reader and reading condition, obtained as the product of between-reader variance σ_b^2 and 1 minus the correlation between the set of AUCs in the two reading conditions, r_b. r_1 = correlation of AUCs within reader, between reading conditions. $r_2 - r_3$ = difference in correlation of AUCs between reader, within reading condition versus between conditions. σ_w^2 = within-reader variance. Variance because cases are a sample, σ_c^2, calculated using a binormal approximation (Obuchowski, 1994)¹²</p>	

4.11. Data Quality Review

Study database. Data for truthing, readings, and lesion matching on pivotal protocol FMSU2017-002B will be provided to BCL following approved data transfer specifications, to be developed by Fujifilm’s study data vendor (Prosoft Clinical, Wayne, PA). *Analysis data: Subject level* will be derived from a subset of protocol FMSU2013-004A *Analysis data: Subject level* in the archive generated by BCL for that study. Fujifilm will transfer reader experience data directly to BCL in comma-separated values (CSV) format.

Review and queries. BCL will examine the database for complete data and, if any data points are missing, query Fujifilm regarding reasons for this missingness. BCL also will verify that data values fall in allowable ranges and follow logical flow, and query Fujifilm regarding any exceptions. Fujifilm will resolve any such issues in the database, and provide responses and an updated database to the statistician. BCL will review the replies and updated database, and declare the data “all clean” if BCL determines that all queries have been resolved sufficiently for analysis to proceed. If data are not “all clean”, BCL will query any remaining exceptions and Fujifilm will reply as above. Data will be locked only after BCL declares the database all clean.

BCL will use this final study database for all final study analysis. Final study analysis may be delayed until the study database is locked.

Missing Responses, Indeterminate Results, and Outliers. BCL will review the reasons for any missing data to evaluate whether the missingness is most likely missing completely at random, missing at random, or systematic. BCL will determine appropriate methods for handling the missing data based on this evaluation and the amount of missingness. If BCL needs to amend this SAP to include more details for handling missing data, we will add these details before carrying out the analysis. In particular, if statistical models are used to address missing data issues these models and their assumptions will be explained clearly, and robustness of results will be explored.

The eCRFs are designed to prevent indeterminate responses; if any do occur, BCL will work with Fujifilm to resolve the issue. Regarding outliers the only continuous variable in the dataset is POM, a subjective ordinal variable for which each reader is permitted to use the full range on each case independent of values of other variables, such that no value of POM in 0 – 100% will be categorized as an outlier.

5. Results to be Reported

- *Dates (timeline).*
 - When cases were accrued on protocol FMSU2013-004A.
 - When cases were selected for the reader study.
 - When the readers' interpretations occurred.
 - When lesion matching occurred.
- *Clinical and demographic characteristics of cases.* For example: age, race, ethnicity, breast composition (BI-RADS breast density categories), study center, reference standard status (cancer, benign, recall, normal); and for cancers lesion type (mass, asymmetry, calcification, architectural distortion, other, or combinations thereof) and size (as determined on protocol FMSU2013-004A).
- *Clinical and demographic characteristics of readers.* For example: years in practice, whether the reader had specialized mammography training, number of mammograms read in the past year, percent of current practice that is mammography, usual hours spent in a clinical day (to address issues of reading fatigue), and whether or not they use C-View.
- *Flow diagram.* Reasons for any exclusions (for example, protocol deviations). If exclusions are minimal, this diagram may be omitted and replaced by text.
- *Summaries and cross-tabulations.*
 - Table of number of findings by reference standard status and modality for each study reader.
 - Means and SDs, and/or medians and quartiles or ranges, of POM requiring correct lesion localization by reference standard status and modality for each study reader.
 - Table of BI-RADS requiring correct lesion localization by reference standard status and modality for each study reader.
 - Table of recall requiring correct lesion localization by reference standard status and modality for each study reader.
- *AUC (primary endpoint).*
 - Graphs of the readers' non-parametric (trapezoidal) ROC curves based on per-subject POM scores requiring correct lesion localization for each review condition (FFDM read, DBT plus S-View read).

- Table of corresponding AUCs for FFDM, DBT plus S-View, and the pairwise differences between them.
- Average across readers of within-modality AUCs and between-modalities differences in AUCs.
- Two-sided 95% CIs to quantify uncertainty in the within-modality estimates and the between-modalities difference.
- The above will be used to evaluate both non-inferiority of AUC (primary endpoint) and superiority of AUC (a secondary endpoint).
- Corresponding rotated forest plots and/or stacked bar charts (optional).
- *Recall rate for all non-cancer cases, recall rate for cancer cases, sensitivity (per-subject, per-lesion), specificity (secondary endpoints).*
 - Table of readers' estimates for FFDM, DBT plus S-View, and the pairwise differences between them.
 - Average across readers of within-modality estimates and between-modalities differences in between them.
 - Two-sided 95% CIs to quantify uncertainty in the within-modality estimates and the between-modalities difference.
 - Results of hypothesis testing using the iterative graphical approach to protect the study's Type 1 error rate from inflation.
 - Corresponding rotated forest plots and/or stacked bar charts (optional).
- *Performance metrics in important subgroups (recall rate in recall, normal, and benign non-cancer cases; lesion-level sensitivity in soft tissue lesions [masses with or without calcifications, focal asymmetries, and/or architectural distortions] and calcifications; any metrics in women with dense breasts and women with non-dense breasts), if analyses in these subgroups are performed.*
 - *AUC only:* Graphs of the readers' non-parametric (trapezoidal) ROC curves based on per-subject POM scores requiring correct lesion localization for each review condition (FFDM read, DBT plus S-View read).
 - Table of readers' estimates for FFDM, DBT plus S-View, and the pairwise differences between them.

- Average across readers of within-modality estimates and between-modalities differences in estimates.
- Two-sided 95% CIs to quantify uncertainty in the within-modality estimates and the between-modalities difference.
- *AUC only*: Corresponding rotated forest plots and/or stacked bar charts (optional).
- *Adverse events*. None are expected; any that are reported to BCL will be described.

6. Regulatory and Administrative Information

If requested, BCL will provide an electronic copy of line data and associated metadata to Fujifilm. Upon regulatory agency request, BCL will provide an electronic copy of statistical software code and/or its output for use in regulatory review, under the conditions of the contract between BCL and Fujifilm.

Analyses will be performed using R version 3.4.1 or later (2017-06-30; R Foundation for Statistical Computing, <https://www.R-project.org>) and cross-validated by standard BCL quality control methods.

7. References

1. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Stat Med.* 2009;28(4):586-604.
2. Lu B, Greevy R, Xu X, Beck C. Optimal nonbipartite matching and its statistical applications. *Am Stat.* 2011;65(1):21–30.
3. U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health (CDRH). Guidance for Industry and FDA Staff: Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data – Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions. July 3, 2012.
4. Wald A, Wolfowitz J. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics.* 1940;11(2):147-62.
5. Campbell G, Pennello G, Yue L. Missing Data in the Regulation of Medical Devices. *J Biopharm Stat.* 2011;21(2):180-195.
6. Obuchowski, NA, Rockette, HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. *Commun Stat Simul Comput.* 1995;24(2):285-308.
7. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat Med.* 2007;26(3):596-619.
8. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data, *Biometrics.* 1992;48(2):577-585.
9. Obuchowski NA. On the comparison of correlated proportions for clustered data. *Stat Med.* 1998;17(13):1495-507.
10. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). *Multiple Endpoints in Clinical Trials – Guidance for Industry* (Draft Guidance). January 2017.
11. Obuchowski, NA. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using analysis of variance with dependent observations. *Acad Radiol.* 1995;2(S1):S22-S29.
12. Obuchowski NA. Computing sample size for receiver operating characteristic studies. *Invest Radiol.* 1994;29(2):238–243.

Appendix 1
Case Report Forms (CRFs)

(CRFs follow)



FUJIFILM Medical Systems U.S.A., Inc.
419 West Avenue
Stamford, Connecticut 06902-6343
1.203.324.2000
FujifilmUSA.com

RADIOLOGIST READER QUESTIONNAIRE – Part 1

Protocol Number(s):	FMSU2017-002B
1. Name:	
2. Are you a radiologist in an academic or community practice?	
3. How many years have you been reading mammograms?	
4. In your most recent MQSA report, how many cases did you review in a year?	
5. How many hours do you read in an average clinic day?	
6. How many years have you been reading FFDM images?	
7. How many years have you been reading DBT images?	
7a. If you read DBT, which Manufacturer do you use?	
7b. What percentage of mammography cases do you perform with DBT?	
7c. If not 100%, what criteria do you use to decide which patients are imaged with DBT?	
7d. Do you use that manufacturer's synthesized view?	
7e. If so, what percentage of mammography cases do you use that manufacturer's synthesized view for?	
8. Any additional comments:	

FMSU2017-002B_Reader Questionnaire_Part 1



FUJIFILM Medical Systems U.S.A., Inc.
419 West Avenue
Stamford, Connecticut 06902-6343
1.203.324.2000
FujifilmUSA.com

RADIOLOGIST READER QUESTIONNAIRE – Part 2

Protocol Number(s):	FMSU2017-002B
1. Name:	
2. With respect to this reader study, what changes to the training program would you suggest?	
3. Please describe your overall impression of the reader study.	
4. Please describe any changes you would suggest to the overall reader study.	
5. Please describe your overall impression of the image quality of the FFDM mammograms.	
5a. Please describe any changes you would suggest to the image quality of the FFDM mammograms.	
6. Please describe your overall impression of the image quality of the DBT images.	
6a. Please describe any changes you would suggest to the image quality of the DBT images.	
7. Please describe your overall impression of the image quality of the S-View images.	
7a. Please describe any changes you would suggest to the image quality of the S-View images.	
8. Please describe your overall impression of the ASPIRE Bellus II Workstation?	
9. Please describe any changes you would suggest to the ASPIRE	

FMSU2017-002B_Reader Questionnaire_Part 2

BIostatistics Consulting, LLC

Bellus II Workstation?	
10. Please describe your overall impression of the travel arrangements/accommodations.	
11. Any additional comments:	

FMSU2017-002B_Reader Questionnaire_Part 2

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2013-004A
EDMS No.: IQG-0031788-D

Subject ID #:

-

Subject Initials (F/M/L):

SUBJECT DEMOGRAPHICS AND INCLUSION/EXCLUSION CRITERIA

1. Date of Consent (mm/dd/yy): - -
2. Age of Subject: 3. Date of Birth (mm/dd/yy): - -
4. Ethnicity: ☐ Hispanic or Latino ☐ Not Hispanic or Latino ☐ Unknown/Not Reported
5. Race: ☐ American Indian or Alaska Native ☐ Black or African American ☐ Asian
☐ Native Hawaiian or other Pacific Islander ☐ White ☐ Other: _____
6. Subject enrolled as (check one): ☐ Screening ☐ Diagnostic ☐ Recall
7. Inclusion Criteria: (Subject may not be enrolled if any criteria are answered as "no")
- | | Yes | No | N/A |
|---|--------------------------|--------------------------|--------------------------|
| a) For the screening-group subjects, is the subject at least 40 years of age, asymptomatic, and scheduled for a routine screening mammogram? | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| b) For the diagnostic-group subjects, is the subject at least 18 years of age; Scheduled for a biopsy due to an assessment of BI-RADS® 4 or 5 after diagnostic work-up of a suspicious screening or clinical finding within the last 60 days. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| c) For the recall-group subjects, is the subject at least 18 years of age; Received a BIRADS 0 within the last 60 days, and are recalled for additional imaging. | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| d) Have the ability to understand the requirements of the study, to provide written informed consent, and to comply with the study protocol | <input type="checkbox"/> | <input type="checkbox"/> | |
| e) Meet none of the exclusion criteria | <input type="checkbox"/> | <input type="checkbox"/> | |

8. Exclusion Criteria: (Subject may not be enrolled if any criteria are answered as "yes")
- | | Yes | No |
|--|--------------------------|--------------------------|
| a) Presence of an Implant | <input type="checkbox"/> | <input type="checkbox"/> |
| b) Women with only a single breast; for example, post mastectomy patients. | <input type="checkbox"/> | <input type="checkbox"/> |
| c) Is pregnant or believes she may be pregnant. | <input type="checkbox"/> | <input type="checkbox"/> |
| d) A woman who has delivered and who has expressed the intention to breast-feed or is currently breast-feeding. | <input type="checkbox"/> | <input type="checkbox"/> |
| e) A woman who has significant existing breast trauma within the last one (1) year. | <input type="checkbox"/> | <input type="checkbox"/> |
| f) Has self-reported severe non-focal or bilateral breast pain affecting subject's ability to tolerate digital mammography and/or breast tomosynthesis examinations. | <input type="checkbox"/> | <input type="checkbox"/> |
| g) A woman who has had a mammogram performed for the purpose of therapy portal planning within the last one (1) year. | <input type="checkbox"/> | <input type="checkbox"/> |
| h) Cannot, for any known reason, undergo follow-up digital mammography and/or breast tomosynthesis examinations (where clinically indicated) at the participating institution. | <input type="checkbox"/> | <input type="checkbox"/> |
| i) Is an inmate (see US Code of Federal Regulations 45CFR46.306) | <input type="checkbox"/> | <input type="checkbox"/> |

Effective Date: 11/12/2014

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2013-004A
EDMS No.: IQG-0031788-D

Subject ID #:

--	--	--	--	--	--

Subject Initials (F/M/L):

--	--	--

MAMMOGRAPHIC FINDINGS

1. Breast Density (check one):

☐

a. Mostly Fatty

☐

b. Scattered Fibroglandular

☐

c. Heterogenously Dense

☐

d. Extremely Dense

2. Routine Mammogram BIRADS Score (check one):

☐

0.

☐

1

☐

2

☐

3

☐

4a

☐

4b

☐

4c

☐

5

3. Does this patient have findings that require work-up? ☐ Yes ☐ No

If YES, please complete a Lesion CRF for each lesion being evaluated.

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Truther Form

Subject ID #:

Date of Reading (mm/dd/yy):

- -

LESION #1

1. Image Type: ☐ FFDM ☐ DBT plus S-View

2. Affected breast: ☐ Right ☐ Left

3. What is the most suspicious finding type? (check all that apply):

a. ☐ Mass

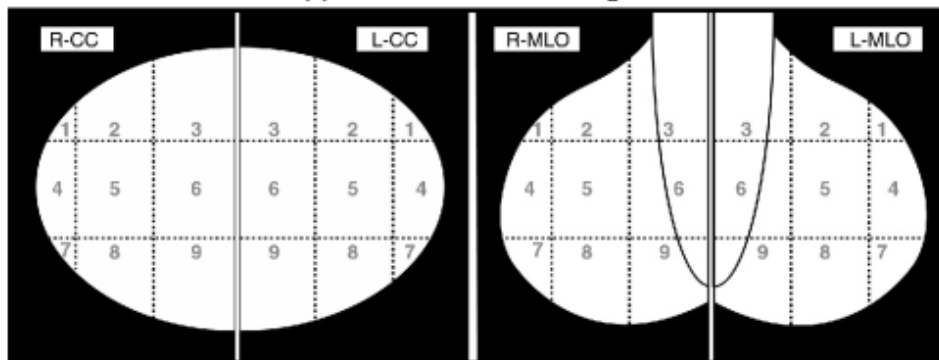
b. ☐ Asymmetry

c. ☐ Calcification

d. ☐ Architectural Distortion

e. ☐ Other (please specify):

4 - 7. Mark an X at the location(s) of the lesion on the Mammogram:



4. 5. 6. 7.

8. Lesion Location: N/A ☐ CC: X Y Slice (DBT Only)

N/A ☐ MLO: X Y Slice (DBT Only)

9. Comments:

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Truther Form

Subject ID #:

Date of Reading (mm/dd/yy):

- -

LESION #2 N/A ☐

10. Affected breast: ☐ Right ☐ Left

11. What is the most suspicious finding type? (check all that apply):

a. ☐ Mass

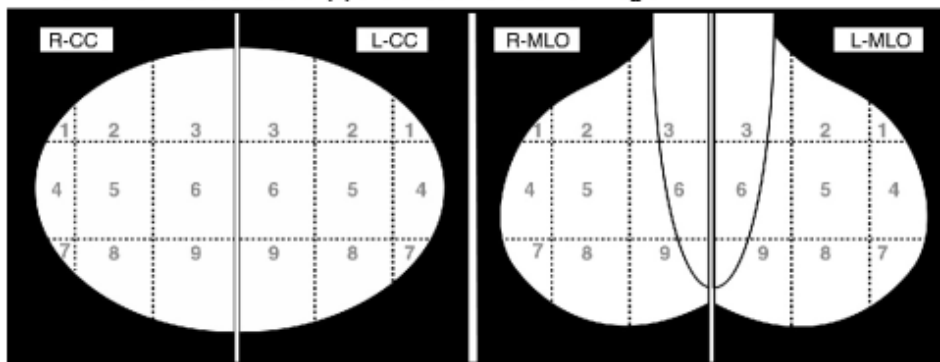
b. ☐ Asymmetry

c. ☐ Calcification

d. ☐ Architectural Distortion

e. ☐ Other (please specify):

12 - 15. Mark an X at the location(s) of the lesion on the Mammogram:



12. 13. 14. 15.

16. Lesion Location: N/A ☐ CC: X Y Slice (DBT Only)

N/A ☐ MLO: X Y Slice (DBT Only)

17. Comments:

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Truther Form

Subject ID #:

Date of Reading (mm/dd/yy):

- -

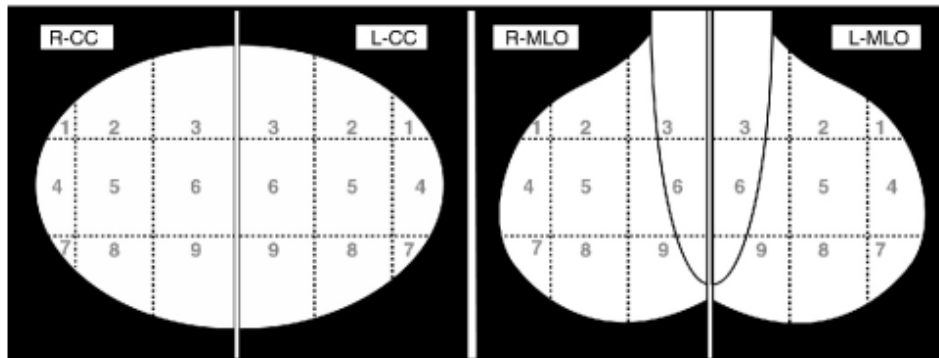
LESION #3 N/A ☐

18. Affected breast: ☐ Right ☐ Left

19. What is the most suspicious finding type? (check all that apply):

- a. ☐ Mass b. ☐ Asymmetry c. ☐ Calcification
 d. ☐ Architectural Distortion e. ☐ Other (please specify):

20 - 23. Mark an X at the location(s) of the lesion on the Mammogram:



20. 21. 22. 23.

24. Lesion Location: N/A ☐ CC: X Y Slice (DBT Only)

N/A ☐ MLO: X Y Slice (DBT Only)

25. Comments:

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Reader Study Form

Sequence ID #:

--	--	--	--	--	--

Date of Reading (mm/dd/yy):

		-			-		
--	--	---	--	--	---	--	--

WRKSTNID #:

--	--	--

INITIAL MAMMOGRAPHIC FINDINGS

1. Are there any mammographic findings in this patient's exam? ☐ Yes ☐ No

2. Initial BI-RADS Score: ☐ 0 ☐ 1 Negative ☐ 2 Benign Finding

3. If BI-RADS 1 or 2, please answer questions 3a and 3b:

3a. Overall, what is your estimated probability of malignancy for this patient (0-100%)?

--	--	--

3b. Would you recall this subject? ☐ Yes ☐ No

If BI-RADS 0, please complete the following questions for up to 3 lesions:

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Reader Study Form

Sequence ID #:

Date of Reading (mm/dd/yy):

- -

WRKSTNID #:

LESION #1

4. Affected breast: ☐ Right ☐ Left

5. What is the most suspicious finding type? (check all that apply):

a. ☐ Mass

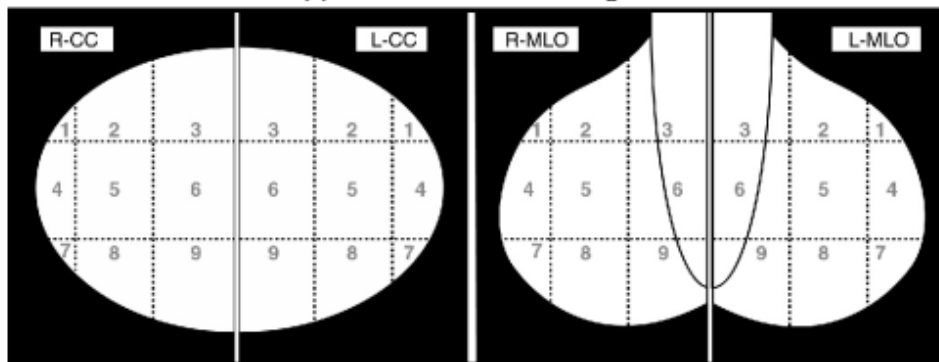
b. ☐ Asymmetry

c. ☐ Calcification

d. ☐ Architectural Distortion

e. ☐ Other (please specify):

6 - 9. Mark an X at the location(s) of the lesion on the Mammogram:



6. 7. 8. 9.

10. Lesion Location: N/A ☐ CC: X Y Slice (DBT Only)

N/A ☐ MLO: X Y Slice (DBT Only)

11. Based on the images reviewed, what is your forced BI-RADS score for this lesion? (check one)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

12. What is your estimated probability of malignancy for this lesion (0-100%)?

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Reader Study Form

Sequence ID #:

Date of Reading (mm/dd/yy):

- -

WRKSTNID #:

LESION #2 N/A ☐

13. Affected breast: ☐ Right ☐ Left

14. What is the most suspicious finding type? (check all that apply):

a. ☐ Mass

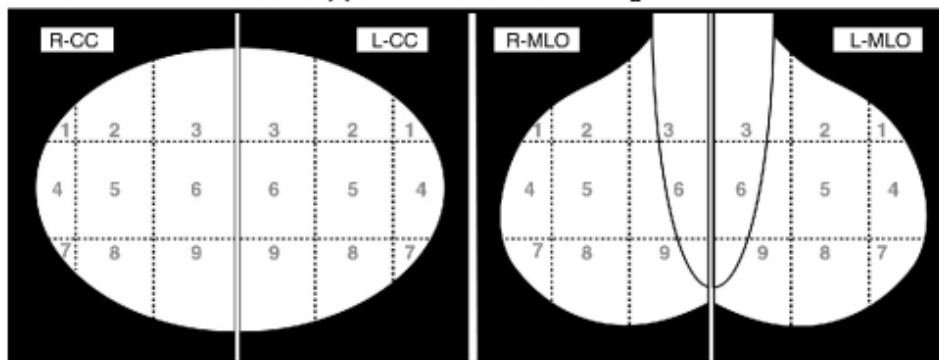
b. ☐ Asymmetry

c. ☐ Calcification

d. ☐ Architectural Distortion

e. ☐ Other (please specify):

15 - 18. Mark an X at the location(s) of the lesion on the Mammogram:



15. 16. 17. 18.

19. Lesion Location: N/A ☐ CC: X Y Slice (DBT Only)

N/A ☐ MLO: X Y Slice (DBT Only)

20. Based on the images reviewed, what is your forced BI-RADS score for this lesion? (check one)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

21. What is your estimated probability of malignancy for this lesion (0-100%)?

VF1.0_09292017
DOC-0036587-A

Page 3
10/9/2017

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Reader Study Form

Sequence ID #:

Date of Reading (mm/dd/yy):

- -

WRKSTNID #:

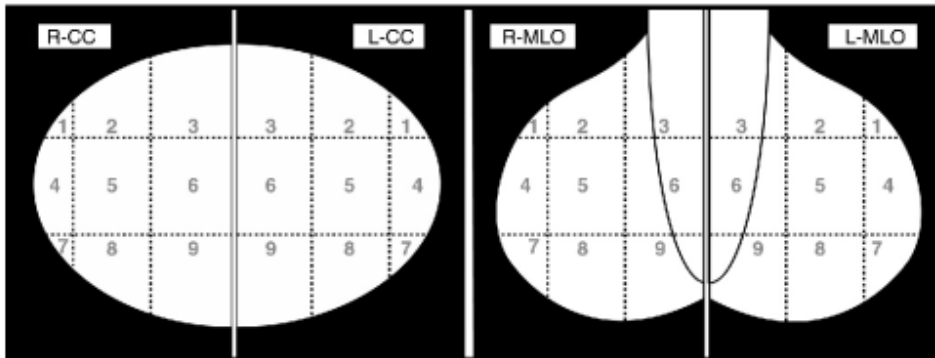
LESION #3 N/A ☐

22. Affected breast: ☐ Right ☐ Left

23. What is the most suspicious finding type? (check all that apply):

- a. ☐ Mass b. ☐ Asymmetry c. ☐ Calcification
d. ☐ Architectural Distortion e. ☐ Other (please specify):

24 - 27. Mark an X at the location(s) of the lesion on the Mammogram:



24. 25. 26. 27.

28. Lesion Location: N/A ☐ CC: X Y Slice (DBT Only)

N/A ☐ MLO: X Y Slice (DBT Only)

29. Based on the images reviewed, what is your forced BI-RADS score for this lesion? (check one)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

30. What is your estimated probability of malignancy for this lesion (0-100%)?

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Reader Study Form

Sequence ID #:

--	--	--	--	--	--

Date of Reading (mm/dd/yy):

		-			-		
--	--	---	--	--	---	--	--

WRKSTNID #:

--	--	--

OVERALL PATIENT RECALL

31. Overall, would you recall this subject? ☐ Yes ☐ No

32. Overall, what is your forced BI-RADS score for this subject? (*check one*)

<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4	<input type="checkbox"/> 5
----------------------------	----------------------------	----------------------------	----------------------------	----------------------------

33. Overall, what is your estimated probability of malignancy for this subject (0-100%)?

--	--	--

FUJIFILM Medical Systems U.S.A., Inc.

PROTOCOL: FMSU2017-002A/B

Lesion Matching Form

Subject ID #:

Date of Review (mm/dd/yy):

1. Image Type: ☐ FFDM ☐ DBT plus S-View

2. Reader Sequence Number:

3. Truther Lesion #1:

Reader Lesion #: ☐ 1 ☐ 2 ☐ 3 ☐ Not Seen

4. Truther Lesion #2 or ☐ NA:

Reader Lesion #: ☐ 1 ☐ 2 ☐ 3 ☐ Not Seen

5. Truther Lesion #3 or ☐ NA:

Reader Lesion #: ☐ 1 ☐ 2 ☐ 3 ☐ Not Seen

6. Comments: