Clinical Study Protocol

Study Title:	Imagio Pivotal Multi-Reader, Multi-Case Study of Optoacoustic Images versus Imagio Ultrasound to Guide Decision to Biopsy
Study Type:	Single arm, sequentially read, controlled, blinded, multi-reader, multi-case study
Study Identifier:	Reader Study-02
Study Phase:	Pivotal
Study Objective:	To prospectively demonstrate Imagio effectiveness for pre-defined endpoints
Indication:	*The Imagio® breast imaging system is indicated for use by a trained and qualified healthcare provider for evaluation of palpable and non-palpable breast abnormalities in women who are referred for a diagnostic breast work-up, following clinical presentation or either screening or diagnostic mammography. The ultrasound mode should be initially used in a targeted fashion, to assess any focal area(s) of clinical or imaging concerns. In ultrasound mode, the device can be used to assign a BI-RADS category to either breast tissue or a mass that is causing clinical or imaging concerns. Masses which are classified as BI-RADS categories 3 through 5 may then be assessed using the Opto-Acoustic (OA) mode. In the OA mode, the Imagio® provides information about the central nidus, boundary and peripheral zones to assist in the diagnosis of the benign or malignant mass(es) of interest. For ultrasound BI-RADS 3-5 masses, using the OA features of the mass allows for improved classification of the mass of interest as compared to ultrasound alone. The OA mode is not indicated for ultrasound BI-RADS 1 and 2 lesions.* This device is not intended to be used as a replacement for mammographic screening or for definitive pathologic diagnosis.
Sponsor:	Seno Medical Instruments, Inc.
Sponsor Contact:	Shaan Schaeffer, Vice President, Clinical Operations Seno Medical Instruments, Inc. 8023 Vantage Dr. Suite 1000 San Antonio, TX 78230 Phone: 210-615-6501 E-mail: sschaeffer@senomedical.com
IDE Approval Identifier	Not Applicable – Non-Significant Risk Study

PROTOCOL SIGNATURE PAGE

Shaan Schaeffer — 5-Nov-2019
Shaan Schaeffer
Vice President of Clinical Operations
Date

A. Thomas Stavros MD — 5-NOV-2019
A. Thomas Stavros, MD
Chief Medical Officer
Date

Ann Waterhouse — 5 Nov 2019
Ann Waterhouse
Vice President of Regulatory and Quality Assurance
Date

Philip Lavin — 5 Nov 2019
Philip Lavin PhD, FASA, FRAPS
Chief Biostatistician
Date

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

ACR	American College of Radiology

AE	Adverse Event
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BBRF	Boston Biostatistics Research Foundation
BI-RADS (BR)	Breast Imaging-Reporting and Data System
CB	Color Balancing
CDU	Conventional Diagnostic Ultrasound
CI	Confidence Interval
CSR	Clinical Study Report
eCRF	Electronic Case Report Form
EDC	Electronic Data Capture System
eTMF	Electronic Trial Master File
FDA	Food and Drug Administration
GCP	Good Clinical Practice
GEE	Generalized Estimating Equations
HIPAA	Health Insurance Portability and Accountability Act
IA	Interim Analysis
ICF	Informed Consent Form
ICH	International Conference on Harmonization
IRB	Institutional Review Board
ITD	Intention-to-Diagnose
IUS	Imagio Internal Ultrasound
LOM	Likelihood of Malignancy
MQSA	Mammography Quality Standards ACR
MRMC	Multi-reader, multi-case
N/A	Not Applicable
NLR	Negative Likelihood Ratio
NPV	Negative Predictive Value
NSR	Non-significant Risk
OA	Opto-Acoustic
OR	Obuchowski-Rockette
PD	Protocol Deviation
PDU	Positive Diagnostic Ultrasound
PDUNB	Positive Diagnostic Ultrasound, No Biopsy
PIONEER	Abbreviated name of a previous study (Pivotal study of Imaging with Opto-acoustics to diagnose breast masses detected by mammography and/or clinical findings: A New Evaluation Tool for Radiologists)
PLR	Positive Likelihood Ratio

PMA	Premarket Approval
POM	Probability of Malignancy
PPV	Positive Predictive Value
QC	Quality Control
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SAP	Statistical Analysis Plan
SAS	Statistical Analysis System
SG	SenoGram
SOP	Standard Operating Procedure
SSRE	Sample Size Re-Estimation
TPB	Truth Panel Benign
TRIAD	Transfer of Images and Data
USA	United States of America

STUDY TEAM LEADERS

Shaan Schaeffer - VP of Clinical Operations

Philip Lavin PhD, FASA, FRAPS - Chief Biostatistician

Tom Stavros MD, FACP - Chief Medical Officer

CLINICAL RESEARCH ORGANIZATIONS

American College of Radiology Center for Research and Innovation (ACR - CRI) will serve as the Imaging Core Lab to conduct the Imagio Pivotal Reader Study. ACR-CRI is located in Philadelphia, PA 19103. ACR-CRI complies with Good Clinical Practices (GCPs) and Title 21 Code of Federal Regulations (CFR) Part 11.

Boston Biostatistics Research Foundation (BBRF) will provide study design, mass sampling, database construction, and data analyses for the Imagio Pivotal Reader Study. BBRF is located in Framingham, MA.

1. STUDY OBJECTIVES

1.1 Primary Objective

To evaluate the following overall and for individual readers for IUS versus Imagio (IUS+ OA):

The gain in *Imagio (Imagio Internal Ultrasound [IUS] + Opto-acoustic [OA]* specificity versus IUS, controlling for sensitivity

1.2 Secondary Objective

To evaluate overall and for individual readers for IUS versus Imagio (IUS+OA): additional effectiveness performance metrics of interest.

2. ENDPOINTS

2.1 Primary Endpoint

To evaluate the following overall and for individual readers for IUS versus Imagio (IUS+OA):

A clinically significant gain of 10% in specificity for 98% sensitivity associated with Imagio (IUS+OA) compared to IUS, with specificities interpolated from the respective ROC curves as required.

2.2 Secondary Endpoints

The following secondary endpoints will be evaluated overall and by individual readers:

- Negative Likelihood Ratio defined as ((1-sensitivity)/specificity)
- Positive Likelihood Ratio defined as (sensitivity/(1-specificity))
- Partial ROC AUC corresponding to 95-100% sensitivity

3. BACKGROUND AND RATIONALE

The Imagio Reader-02 Pivotal Study is intended to evaluate if the results observed in the previous Reader-01 Feasibility Study can be confirmed for pre-specified effectiveness endpoints. Using Intention-to-Diagnose (ITD) masses from the PIONEER Pivotal Study, 480

to 840 masses are to be used for the Reader-02 Pivotal Study, the remaining masses will be reserved for training the SenoGram tool.

4. SENOGRAM

The SenoGram is a tool that helps the radiologist predict the likelihood of malignancy (LOM) based on a set of reader-assigned feature scores and other relevant data. The SenoGram classification model was validated and locked prior to the Reader-01 Feasibility Study. The algorithm, feature set, and development method remained locked for the Reader-02 Pivotal study. The SenoGram model will be refit to a subset of the PIONEER dataset that excludes all masses in the Reader-02 Pivotal study. The refit model will be locked prior to the beginning of this Reader-02 study. For this Reader-02 study, the SenoGram tool will consist of two components, a user interface and a back-end computation engine.

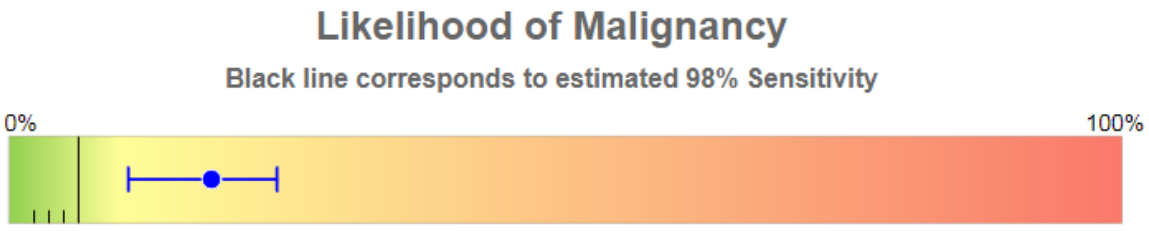
The reader will enter 5 feature scores for IUS, 5 feature scores for OA and 4 other features: subject age, maximum diameter depth to posterior wall of mass, and mammographic BI-RADS category (when available). When all data has been entered the user interface will graphically display the SenoGram LOM.

The reader will then enter his/her final assessments in the following order after viewing the SenoGram LOM based on his/her knowledge from experience and IUS+OA training:

1. Probability of Malignancy (POM) and
2. Subsequent BI-RADs category

The SenoGram output figure is shown below. The SenoGram back-end uses an ensemble of models to compute a set of predicted probabilities. The mean of these predicted probabilities is indicated by the blue circle, and the 5% to 95% prediction interval is depicted as a blue horizontal line through the circle. For reference purposes the SenoGram prediction is shown relative to the suggested biopsy/no-biopsy threshold, depicted by a black vertical line on the left side of the bar. This threshold corresponds to an estimated 2% false negative rate (i.e., 98% sensitivity). The SenoGram uses machine learning to assist the radiologist in assimilating and weighting 14 features that impact the LOM.

Figure 1: Likelihood of Malignancy



5. STUDY DESIGN

This will be a single arm, controlled, blinded, multi-reader, multi-case (MRMC) study using a sequential design. The study will include 15 readers with an additional 5 back-up readers. Readers that participated in the Seno PIONEER Study, the MAESTRO Study, or the Reader-01 Feasibility Study are not eligible to participate as readers in this Reader-02 Pivotal Study.

Imagio (IUS+OA) training will be completed prior to any reads taking place. Read 1 will be immediately followed by Read 2 within the same read session.

- Read 1 (Control): History + Mammogram (if available) + IUS (stills and videos provided), IUS Probability of Malignancy (POM) and BI-RADS category assigned in the data form then locked.

- Read 2 (Test): History + Mammogram (if available) + IUS (stills and videos provided), and Imagio (IUS+OA) (stills and videos provided). Imagio (IUS+OA) POM and BI-RADS category assigned after viewing the SenoGram (SG) output. The data form is then locked.

Read 1 reflects the typical information available to a radiologist when evaluating standard ultrasound images, taking into consideration the mass, patient history and assessing mammogram BI-RADS results, when available. The History details provided are summarized under the Blinding section of the protocol. Read 1 (IUS reads) will serve as the control.

Read 2 will display the IUS + OA images. Each reader will record IUS and OA feature scores within the case report form and the SenoGram will display a prediction interval based upon the reader feature score input. Each reader will first assign final POM and then a BI-RADS score. At this point, the Read 2 form is locked.

The Reader-02 Pivotal Study will consist of between 480 to 840 masses with complete imaging read sets from the original PIONEER Intention-to Diagnose (ITD)/analysis population. The masses will be randomly selected from within the PIONEER Study. The data will be organized in blocks of 120 masses, each consisting of 72 benign plus 3 high risk (to be categorized as benign) and 45 malignant masses (reflecting the same 37.5% prevalence of cancer as the overall PIONEER ITD/analysis population). The masses were classified by conventional diagnostic ultrasound (CDU) as BI-RADS 3 to 5 and will be selected at random in proportion to the original assignment distribution of BI-RADS classifications among subjects in the PIONEER Study. To facilitate the alignment of the PIONEER Pivotal Study data with the Reader 02 Pivotal Study data, the mass image set sampling plan will select a benign mass proportion with and without mammograms depending on availability of the mammograms; this stratification will not apply for malignant masses where nearly all masses were previously evaluated using mammography. Blocks of 120 masses will be used for the effectiveness endpoint analyses with a blinded interim analysis conducted after all readers evaluate 360 (360-10 color balance masses= 350) masses with sample size re-estimation used to possibly increase the sample size up to 840 masses depending on the underlying reader variability driving sample size for the primary effectiveness endpoint. A total of 29 image sets used to tune the color balance algorithm (among the 840 masses) will be replaced with 29 comparable image sets in accordance with the Reader-02 Mass Sampling Plan for purposes of the final analysis; these replacement masses will come from the set of 60 overage masses from the PIONEER Pivotal Study plus masses from the PIONEER Pilot Study if there are insufficient overage cases.

6. SELECTION OF READERS

6.1 Reader Qualification Criteria

- Completed residency and are board certified in radiology
- Active breast imager for at least 3 years.
- Readers to meet mammography interpretation requirements per Mammography Quality Standards ACR (MQSA) for the year prior to study
- Readers to meet breast imaging ultrasound interpretation requirements per ACR for year prior to study
- For the clinical study, a willingness to use BI-RADS 4 sub-categories

- For the clinical study, the ability to participate and read all masses in both IUS and IUS+OA reader sessions

6.2 Reader Training

Imagio (IUS+OA) Training Summary Reader 02 Pivotal Study

1. Didactic training module
   a) Fundamentals of OA *(Questions and Answers)*
   b) OA feature scoring *(Questions and Answers)*
   c) Brief summary of Seno learnings from previous studies
   d) OA-histologic correlation and False Positive cases *(Questions and Answers)*
   e) OA artifacts *(Questions and Answers)*
   f) IUS feature scoring, including importance of BI-RADS 4 subcategories, NLR and Bayes Theorem *(Questions and Answers)*
   g) PIONEER Pivotal Study False Negative analysis *(Questions and Answers)*
   h) The SenoGram and how to use It
   i) Summary *(final questions and comments)*
2. Interactive reading training case module (up to 30 cases)
   a) Mixture of benign and malignant cases enriched to 50% malignant cases
   b) Mixture of cases with good, average, and below average PIONEER reader performance – will start with easy cases, move to average cases, and finish with most difficult cases.
   c) Readers will learn to use a reading station, draw regions of interest (ROIs), score IUS and OA features, use the SenoGram to aid in predicting OA POM and BI-RADS category
   d) Each case will be read and scored by the reader trainees. The Seno instructor will review how and why he/she would score the case. Case histology will be reviewed and discussed with concordance or discordance of OA feature scoring.
3. Test Cases - 30
   a) Mixture of benign (18) and malignant (12) masses, same prevalence as overall PIONEER ITD population
   b) Mixture of easy, average, and difficult cases based upon PIONEER Pivotal Study reads
4. Pass / Fail criteria

   Readers reading must pass a proficiency test involving the scoring and interpretation of 30 cases before starting their study reads. If this is not achieved the first time, then the reader takes the test a second time. The reader will be given remediation training targeted to the masses for which they made errors within the first 30 case test set, and the reader will be given a second opportunity to take the test on a different set of 30 test cases. If a reader fails the second test, remediation will take place before they start the Reader-02 Pivotal Study reads. The reader will proceed to read Reader-02 pivotal cases whether or not they pass the second test

7. SELECTION OF READER SETS

7.1 Inclusion Criteria

Reader sets must meet all the following inclusion criteria to be included in this study:

- One analyzable mass per patient: BI-RADS 3, 4a, 4b, 4c, and 5 masses as declared by clinical site investigator via PIONEER study inclusion criteria and categorized as BIRADS 3, 4a, 4b, 4c, and 5 by conventional diagnostic ultrasound (CDU)
- Masses declared to be in the PIONEER ITD/analysis population, including high risk cases per original PIONEER protocol
- Patient age, indication for study entry and available medical history
- Evaluable mammograms (when available) and IUS and OA video loops and still images for each mass

7.2 Exclusion Criteria

Reader image sets must be excluded if any of the following criteria are met:

- Critical missing IUS or OA still image and/or video loop views or incorrect IUS or OA stills and video loops that would preclude a case from being evaluated by readers
- Reader-02 Proficiency Test and training cases

7.3 Reader Set Selection Procedure

- BBRF will develop and apply the Mass Sampling Plan which takes training cases and stratification requirements into consideration
- BBRF will prepare the mass read order per block to be the same for all readers
- All image sets will be de-identified

8. READER STUDY PROCEDURES

8.1 Reader Study Process Work Flow

Figure 2 below illustrates, at a high level, the central reader study process flow. Seno Medical will electronically transfer image sets to ACR-CRI. Upon receipt of the images at ACR CRI, imaging support staff will upload images to **Tr**ansfer of **I**mages and **D**ata [TRIAD], ACR's electronic image submission tool.

The images automatically undergo a de-identification process in TRIAD™, whereby all personal identifier DICOM tags in the image metadata are de-identified according to TRIAD™'s anonymization profile. If any personal identifiers are burned into exams received, ACR CRI staff will remove that by pixel cleaning. ACR-CRI imaging technologists will first perform quality control (QC-1) to document an inventory of all exams and their attributes.

ACR-CRI's contracted multiple mass quality control reviewer will assess the image sets (QC-2) for inclusion and exclusion as well as identifying and labeling Mass 1 on all modalities submitted for a given subject case. The acceptable image sets from the ACR QC (QC-1 and QC-2) process will be incorporated by BBRF and considered ready for central read.

BBRF is solely responsible for the selection of all masses and will generate the randomized study mass list in accordance with the Mass Sampling Plan. The order of masses per block will be the same for each participating reader.

The QC-1 and QC-2 checks will be documented in study specific procedure documents and outputs of such checks documented on study specific forms, all of which will be archived in an electronic Trial Master File (eTMF) and ACR CRI clinical databases.

ACR research staff serving as read monitors will be present during read sessions to ensure the read sequence is maintained, eCRFs are fully completed by the reader, cross-check the assigned read list with completed eCRFs and assist in any workstation-related questions. Each reader will be separately proctored by read monitors to ensure that no technical issues arise, and that reader scoring is independent. The read monitors will relay any issues recognized from the readers or monitors they cannot resolve to the ACR Project Manager for remediation.

A manual process of data entry will be performed. In order to maximize both workflow efficiency and data integrity, this process will be achieved by using the combination of the reader and a read monitor, working together as a team. Both the reader and the read monitor will be trained on the data entry process prior to production subject case assessments. The process of data entry is outlined below.

1. Read monitor calls out the Read ID to be reviewed based on the provided Reader Work List. The reader opens the Subject image case in the image viewing workstation. The read monitor opens the Subject eCRF case in the Rave Electronic Data Capture (EDC) Database. The reader calls out the Subject ID for the opened case as confirmation that it corresponds to the Subject ID opened in the eCRF database by the read monitor.
2. Reader performs the specified subject's image review using the designated Workstation and following assessment criteria as trained on for this study.
3. Read monitor prepares to perform data entry for the specified subject.
4. When the reader is ready to provide the data for data entry, the reader verbally notifies the read monitor.
5. The reader verbally provides the answers to each of the eCRF required field entries.
6. As the read monitor performs the manual data entry for each of the sequential fields directed by the flow of the eCRF, the read monitor verbally calls out their entry which is then verbally confirmed (or corrected, as necessary) by the reader.
7. This process will be repeated until all the appropriate fields in the eCRF for each of the subject cases are completed.

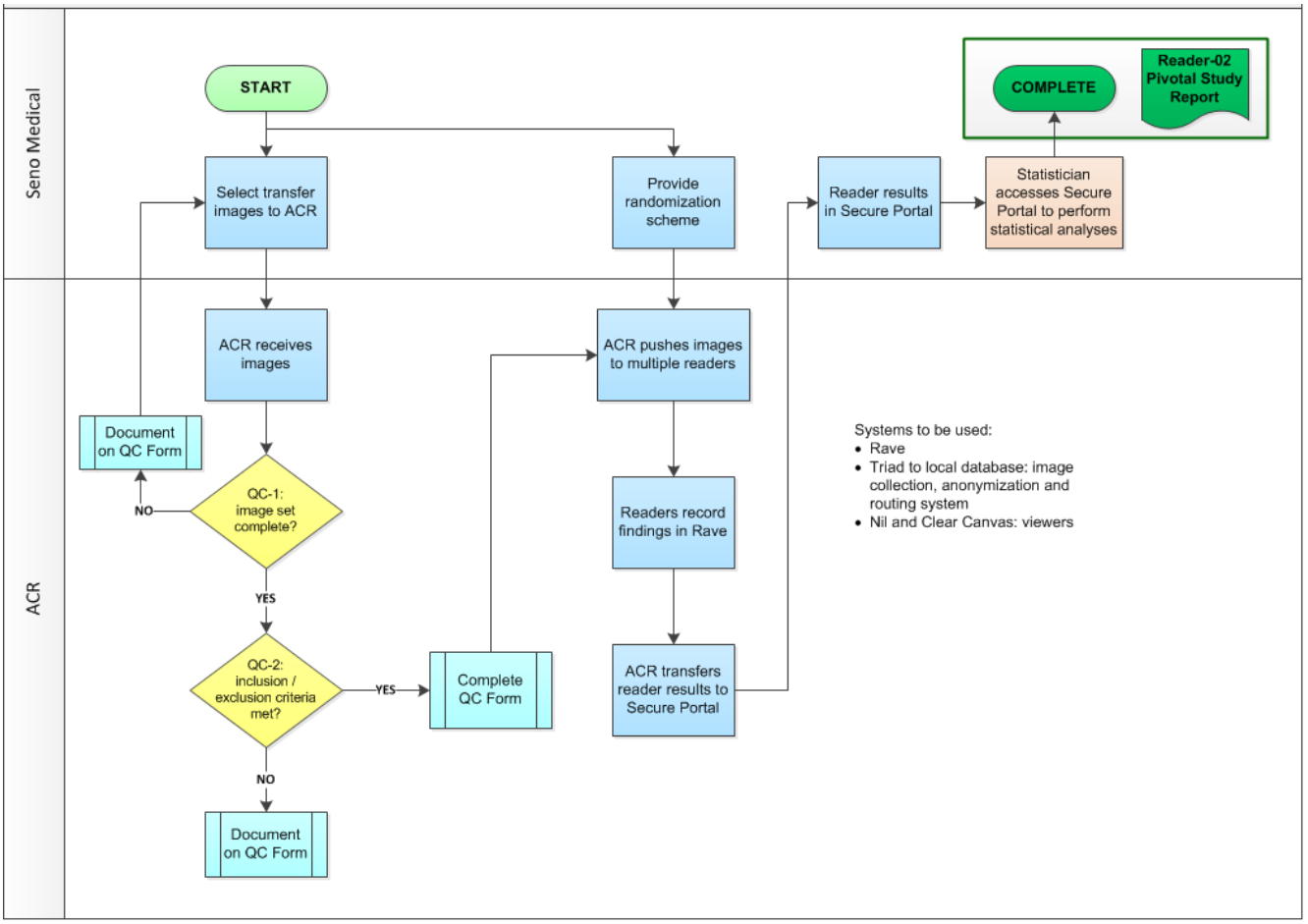
The reader electronically signs off on the data for a subject case at the completion of his/her review. It is understood that the commitment of a reader's electronic signature on the eCRF signifies that the assessment of the corresponding randomization ID is complete and accurate and is an attestation that no changes/edits need to occur.

The read monitor should be attuned to reader performance and suggest breaks as necessary to avoid reader fatigue.

As per the image review and data entry workflow described in this Protocol, once a subject scan review is completed by a reader, the reader provides their electronic signature and the eCRF is 'locked', preventing further changes.

Upon completion of all central read assessments by all readers, ACR-CRI will post the results to ACR's secure portal, BBRF will have access to the portal for purposes of conducting the statistical analyses following database locks for the interim and final analyses.

Figure 2; Reader-02 Pivotal Study Work Flow



8.2 Schedule of Reader Sessions

All reads will include standard reader training on an image viewing workstation including a mammography monitor as well as training on eCRF completion within the Medidata RAVE Electronic Data Capture (EDC) system, etc.

Imagio (IUS+OA) training, which includes training on the SenoGram, and a proficiency test will take place prior to the start of all study reads.

Readers read blocks including IUS and Imagio (IUS+OA) reads in sequential order in accordance to their schedule and read speed.

8.3 Description of Reading Environment

After completion of standard training and Imagio (IUS+OA) training, each of the readers will use an image display and electronic data capture workstation to perform image analysis, working alongside an ACR read monitor, for all reading sessions. Each reader and read monitor will be in a designated, private reading room for their use during a reading session. Readers will work undistracted so that their attention is focused on an accurate interpretation. The workstation desk will have adjustable height for reader comfort, the room will be equipped with noise abatement features, diagnostic display monitors and moderate illumination. As part of reader training and as documented in the confidentiality agreement they sign during the ACR CRI contracting process, readers will be aware and monitored to maintain confidentiality

on study details and subject images reviewed. Access to the reading rooms is controlled and read monitors ensure readers do not take any unauthorized material into the reading rooms.

8.4 Randomization

BBRF will generate the randomly selected read list. The read order will be the same for all readers within each block of 120 masses presented.

8.5 Blinding

The identification of cases randomly selected from the PIONEER dataset to include in the study, as well as the original reader results for those cases, will be blinded to all Seno personnel except individuals directly involved in the study. The assigned unblinded statistician responsible for performing the planned interim analysis will be the only individual to have access to the interim Reader-02 Pivotal Study analysis results.

The readers will be blinded to the diagnosis, but the reader will have access to the following baseline variables:

- Age
- Indication (reason study was ordered)
- Mass location
- Medical History as available
- Mammograms (if available).

Readers will also not have access to results from other readers or to the original biopsy or follow-up outcomes during the study.

9. SAFETY ASSESSMENTS

Safety assessments are not required as this is a reader study of de-identified Imagio (IUS+OA) and IUS images. No patient diagnoses are affected by the reader sets or this MRMC study.

10. STATISTICAL CONSIDERATIONS

10.1 General Design Issues

BBRF will implement the Reader-02 Pivotal Study Mass Sampling Plan to randomly select masses.

The unit of measurement for reader-based outcomes is at the reader-mass level. The Reader-01 Feasibility Study demonstrated that readers could conduct POM evaluation separately for IUS and then for Imagio (IUS+OA) (primary) to evaluate the primary and secondary effectiveness endpoints.

The primary endpoint is to show a specificity advantage for fixed 98% sensitivity (OR-DBM); this diffuses the tradeoff between FNs and FPs if the FN rate is the same. The sensitivities and specificities will be calculated overall and per reader using generalized estimating equations (GEE) and the analysis of variance (ANOVA) modeling as well as for observed without adjustment.

For each primary and secondary endpoint, Imagio (IUS +OA) data will be compared to IUS data, with the analysis stratified by the availability/non-availability of mammogram images.

The specificity advantages will be computed for fixed sensitivity (e.g., 98% target) with interpolation used by calculating the specificities and sensitivities from the estimated ROC

curve as needed. This same principle will be used to construct the partial ROC curves and to compute the area under the ROC curve over 95-100% sensitivity. Specificity at fixed sensitivity and partial AUC will be determined from parametric ROC curves constructed using techniques such as binormal or PROPROC models.

The supporting statistical methodology (2-8) relies on publications by Hillis, Rockette, Obuchowski, and Metz in accordance with best regulatory statistics practice. Readers will be treated as correlated and as random effects.

The Reader-01 Feasibility Study results were used to determine the Reader-02 Pivotal Study sample size. The overall difference in specificity for fixed 98% sensitivity will be assessed to judge if the Reader-02 Pivotal Study can be conducted with 80% power to detect a 10% absolute advantage in specificity (alternative hypothesis) versus no absolute advantage (null hypothesis) with two-sided 5% Type I error where both readers and masses (image sets) are considered to be random effects; this approach leads to the most conservative test given that it is based on the largest source of inter-reader variability. Reader-01 Feasibility Study data will not be used or combined with any subsequent Reader-02 Pivotal Study data.

A blinded interim analysis is planned after all readers complete and record their findings for 360 (360-10 color balance masses=350) masses (first three blocks) with the final analysis to be conducted after 480 to 840 image sets are read; the final sample size will depend upon the intra- and inter-reader variability. Ten color balanced masses will be excluded from the interim analysis. The interim analysis will assess the intra- and inter-reader variabilities in order to determine the final sample size (up to 840 masses). This is not an adaptive study; under no circumstances will the null hypothesis, the alternative hypothesis, population of masses, or choice of endpoint be modified as a result of the interim analysis.

10.2 Reader / Image Set Sample Size

The sample size will be based on the primary effectiveness endpoint, namely the gain in specificity for fixed 98% sensitivity. This calculation depends upon the number of image sets (one per mass) to be read, the number of readers, as well as the inter-reader variability associated with the underlying models used for analysis.

The test of the primary hypothesis is powered to detect a 10% absolute specificity advantage (superiority) as the alternative hypothesis with no specificity advantage as the null hypothesis. The primary endpoint will be evaluated in the following testable hypotheses:

H_0: $S_{Imagio} = S_{IUS}$

H_1: $S_{Imagio} \neq S_{IUS}$

where S_{Imagio} and S_{IUS} represent specificity, values associated with Imagio (IUS + OA) and IUS.

The sample size calculations conservatively assume that both readers and masses are random effects; an ANOVA model will be used to compute sample size. The sample size was computed using the following statistical software written by Stephen Hillis and entitled: Multi-Reader Sample Size Program for Diagnostic Studies [http://perception.radiology.uiowa.edu.].

For 80% power, a two-sided hypothesis test with a two-sided 5% alpha, and 15 readers, a sample size of 854 image sets is needed to detect a 10% absolute advantage in specificity for Imagio (IUS + OA) vs IUS, assuming a fixed 98% sensitivity and the same intra- and inter-reader standard deviation observed in the Reader-1 Feasibility Study.

The supporting calculations including the software output follow:

Results for Feasibility variance estimates

User-supplied parameter or pilot-study values:

```
            Design : factorial
             Tests : nonequivalence
   Readers and cases : both readers and cases random
       Input values : treat as known
             Alpha : 0.05
       Input Format : OR variance components with error covariances
test*reader var comp : 0.00203538
      Error variance : 0.03571
                Cov1 : 0.01106919
                Cov2 : 0.0061484
                Cov3 : 0.0043063
                  c* : 120
```

User-supplied desired power, proposed readers & cases values

```
      Desired Power : 0.8
Proposed max readers : 15
  Proposed max cases : 2000
Proposed min readers : 15
  Proposed min cases : 20
```

Corresponding OR variance components, covariances, and correlations

```
test*reader var comp : 0.00203538
      Error variance : 0.03571
                Cov1 : 0.01106919
                Cov2 : 0.0061484
                Cov3 : 0.0043063
                  r1 : 0.309974517
                  r2 : 0.172175861
                  r3 : 0.120590871
```

Sample Size Results

effect size	readers	cases	power
0.1	15	854	0.801

Initially, 360 masses will be read by all readers for the purposes of performing the interim analysis to assess the sources of variation and to rerun the sample size estimate for the primary effectiveness endpoint. Neither futility nor efficacy will be assessed; this is not an adaptive design.

The sample size may be increased from the minimum of 480 image sets to a maximum of 840 image sets (7 complete blocks) in increments of 120 masses per block. All original color balance masses will be replaced with comparable masses and image sets within the final analysis from the PIONEER Pivotal overage masses.

The maximum sample size to be used for the study will be capped at 840 masses. An overage mass set will be set aside in case quality control issues are found with the images and replacement masses are needed. After excluding the pool of 840 masses and overage masses, approximately 900 masses will remain for independent SenoGram training. No masses used to train the SenoGram, will be in the pool of 840 study masses plus overage. We have removed the 29 masses from the pool of study masses which were mistakenly included from the original color balance algorithm.

Every effort will be made to identify missing reader generated data in real time. Each block of 120 masses will include 75 benign (72 benign + 3 high risk) and 45 malignant masses to represent the PIONEER study.

With the exception of masses excluded due to protocol deviations, all complete reads will be included in all analyses, consistent with an “Intention-to-Diagnose” approach.

10.3 Additional Hypothesis Tests

The following two-sided hypothesis tests will be applied using the observed estimates for NLR and PLR as well as using the PROPROC for the partial ROC AUC:

- NLR:
  - H_0: $NLR_{IUS} = NLR_{Imagio}$ vs
  - H_A: $NLR_{IUS} \neq NLR_{Imagio}$ representing a reduction
- PLR:
  - H_0: $PLR_{IUS} = PLR_{Imagio}$ vs
  - H_A: $PLR_{IUS} \neq PLR_{Imagio}$ representing an increase
- Partial ROC AUC:
  - H_0: $pAUC_{IUS} = pAUC_{Imagio}$ vs
  - H_A: $pAUC_{IUS} \neq pAUC_{Imagio}$ representing an increase

Hierarchical testing will be applied to seek labeling claims if warranted by the pre-defined testing procedure (see Section 10.8).

10.4 Population Definition

Masses and images will be selected for use in accordance with the study protocol, inclusion and exclusion criteria.

10.5 Interim Analysis

10.5.1 Plan

The interim analysis (IA) is to re-estimate sample size to avoid running an underpowered study. The IA will be performed by an unblinded statistician. The Sponsor and all other parties will remain blinded to the results. The only communication to the sponsor will be the final sample size recommendation. The sample size may be increased up to 840 masses from the initial 480 masses in increments of 120 masses per block in order to achieve 80% power for the primary endpoint. This recommendation will be based solely on variance parameters re-estimated in the IA (but not the effect size). Thus, these IA results fall into the category of non-comparative results. Thus, the IA has negligible effect on the Type 1 error. The IA will be based on the total variance estimate (inclusive of both reader and imaging modality variations) so, the IA remains 'non-comparative' since the interim treatment effect was not used to compute the final sample size.

10.5.2 Methodology

The sole IA goal is to compute the final sample size based on the primary endpoint. The sample size re-estimation (SSRE) will be performed without any adaptation. SSRE will be conducted to detect the pre-planned 10% absolute advantage in specificity for fixed 98% sensitivity, given that this is an established criterion agreed upon with FDA. The standard deviation (from reader as a random effect) is not known. There will be no preplanned or performed assessments for futility, stopping for early efficacy, dropping readers, altering the pre-planned 10% effect hypothesis, or any other adaptation. Therefore, this is classic SSRE.

Huang and Chen (1) references this application as distinct from their discussion of adaptive designs. Our IA is based entirely on non-comparative data, that is, no information about treatment effects were incorporated into the decision process. For this SSRE situation (a non-comparative interim), it can be shown that performing analyses at the conventional one-sided 0.025 significance level has a negligible effect on the Type 1 error probability (FDA Guidance

on Adaptive Trials (2), Wittes (3), Kieser and Friede (4), Friede and Kieser (5), among others). Wittes advises that the alpha spend is small when the observed variance is less than the assumed variance (as per the interim analysis) without accounting for imaging modality. Huang and Chen advise that the alpha spend is small when the conditional power exceeds 60% which is the case when a 10% delta is assumed; any delta effect would further reduce the reader variance. Thus, the alpha spend for conducting the planned interim analysis is small.

From a testing perspective, the conclusions should hold in recognition that the p-value is a Z-score whether derived from a one-sample or a two-sample test independent of the model used (Huang or Hillis). All alpha adjustment depends on a linear combination of Z-scores (6) for respective Z-scores prior to and after the pre-planned interim analysis. Thus, first principles all lead to the conclusion that the alpha spend is small enough to ignore. Our estimate is that the alpha spend is no more than 0.01% under the given conditions. The minimal alpha spend derives from the abundant two-sample literature to apply to our paired sample setting with no published literature.

10.5.3 Timing

The single blinded interim analysis is planned once 360 (360-10 color balance masses=350) completed image sets are read by each reader and the data become available. The interim analysis will be performed by an independent, unblinded third party statistician who will make recommendations below to Seno Medical Inc. based on the results from this initial cohort of masses:

1. Sample Size Calculation:
   ___One additional block is required (480 masses) OR
   ___Two additional blocks are required (600 masses) OR
   ___Three additional blocks are required (720 masses) OR
   ___Four additional blocks are required (840 masses) OR
   ___Four additional blocks (840 masses) are not sufficient but reads should continue because there is a reasonable chance to reach statistical significance based on the conditional power assuming the alternative hypothesis holds.
   ___*Four additional blocks (840 masses) are not sufficient, reads should not continue because* there is not a reasonable chance to reach statistical significance based on the conditional power assuming the alternative hypothesis holds.

To maintain blinding to the interim results, no specific results from the interim analysis will be shared by the unblinded party performing the analysis with Seno or any other study participant, and only one of the above recommendations will be provided. Seno retains the right to accept, not accept, or modify any recommendation concerning changes to conduct of the study.

There are no plans to stop the study early for benefit or a positive outcome on the basis of the interim analysis results. No alpha will be spent to perform this interim analysis since this analysis is for sample size re-estimation only (1). Under no circumstances will the alternative hypothesis be modified as a result of this analysis. The final analysis will include the results from the n=360 (360-10 color balance masses=350) interim analysis.

10.5.4 Execution

The unblinded biostatistician will only evaluate the overall SD which was the nuisance parameter driving sample size for the fixed 10% difference to be detected. The unblinded biostatistician will not need to take the observed delta into account. Other than the preplanned sample size calculation to determine the remaining blocks required to complete the study, no

other information will be revealed to the Sponsor or to any other members of the analysis team. The IA will be executed to the blinded principle.

10.6 Protocol Deviations

The protocol deviations will be reviewed in advance of the interim and final analyses for any unreadable, unevaluable, or skipped image sets. A Protocol Deviation Meeting will be held among results-blinded parties prior to the preplanned n=360 interim analysis and final analysis to address how to handle impacted image sets.

At study completion, the number and type of protocol deviations will be presented overall and by reader to determine whether there are statistical concerns. The biostatistician will report findings to the study sponsor.

10.7 Outcomes

Outcomes will be analyzed according to the PIONEER Truth Panel findings and the pathology diagnosis (if available) as truth; Truth Panel Benign (TPB) will be a benign diagnosis. The final analyses will be performed following database lock.

10.8 Hierarchical Testing

To extend labeling claims beyond the primary endpoint, a hierarchical testing strategy will be used to control Type 1 error associated with testing the primary endpoint and three secondary endpoints in the following pre-defined order until statistical significance is no longer reached in support of extended labeling:

- Increase in specificity for fixed 98% sensitivity.
- NLR
- PLR
- pAUC over 95% to 100% sensitivity

No other endpoints are to be included as labeling claims.

10.9 Data Analyses

All analyses will be performed in a GCP-controlled environment using the Hillis software (see references 3-7), SAS v9.3 or higher, and StatXact v10 or later unless otherwise specified.

See the companion SAP for further details.

All significance testing will be two-sided; results will be presented using two-sided p-values and two-sided 95% confidence intervals.

All final analyses will exclude the 29 Color Balanced masses; 350 masses will be analyzed at the interim analysis timepoint. This excludes the 10 Color Balance masses. For the final analysis, the 29 Color Balanced masses will be replaced from the overage and the PIONEER Pilot Study if there are insufficient masses among the overage masses.

Methods of analysis will be based on MRMC analysis methods by Obuchowski and Rockette where both reader and mass are random effects (see references 7-13). Two-sided 95% confidence intervals will be computed for the individual outcomes (IUS, Imagio (IUS+OA)) and paired differences.

The same approach will be used to compute the differences in partial ROC AUC and full ROC

AUC, sensitivities, specificities, and paired differences, for benign and malignant masses. These analyses will be performed for IUS versus Imagio (IUS+OA) SG).

For each endpoint, subgroup analyses will be performed in the same manner for those with a prior mammogram as well as for those without a prior mammogram.

1. Sensitivity, Specificity, Partial ROC AUC, Full ROC AUC, and Specificity for Fixed Sensitivity

Individual reader specificity and sensitivity will be calculated from simple counts for the 2% POMs to construct the respective ROC curves for IUS, the Imagio (IUS+OA).

For sensitivity and specificity, GEE and ANOVA will be used to compute differences as well as to compute the two-sided 95% confidence intervals for the OA-IUS differences.

The specificity advantage for fixed 98% sensitivity will be the primary endpoint; a 10% absolute advantage for Imagio vs IUS in specificity for fixed 98% sensitivity is considered to be clinically relevant; this drives the final sample size calculation. The Hillis software will be used to analyze the primary endpoint as well as for the partial and full ROC AUCs. The AUCs will be computed from the respective ROC curves which are generated according to the POMs.

Specificity and sensitivity are calculated using this fixed 2% cutoff. Please be assured that the OR-DBM MRMC 2.51 (Hillis) software does not pick or use a fixed cutoff. In our application, the Hillis software computes AUC, partial AUC, specificity gain at fixed sensitivity, and sensitivity gain at fixed specificity using a full factorial ANOVA incorporating terms for readers, masses, and modalities. The gain in specificity can be computed at a wide range of sensitivities. The 98% sensitivity corresponds to a 2% false negative (FN) rate, but this unrelated to the use of a fixed 2% POM cutoff for clinical decision making.

The partial AUC will be computed over the 95-100% sensitivity range using PROPROC. To compute the specificity advantage for fixed sensitivity, interpolation may be required; a fixed sensitivity will be targeted, e.g., 98%. To compute the specificity advantage and the partial area under the ROC, the specificities and sensitivities may need to be further computed from the ROC curves as needed. The variances will be computed to construct the two-sided 95% CIs for Imagio (IUS+OA), IUS, and for the respective pairwise Imagio (IUS+OA) - IUS difference.

2. NLR and PLR

NLR and PLR will be analyzed using the logarithmic transformation and the delta method. Calculations will be performed using MATLAB script: DLRATIOS.M. The method of Nofuentes (14) will be used; this reflects that the NLR and PLR are both ratios of independent random variables. The DLRATIOS routine computes the standard deviation for the ratios of two independent random variables corresponding to the sensitivity and specificity which are both used to compute NLR and DLR.

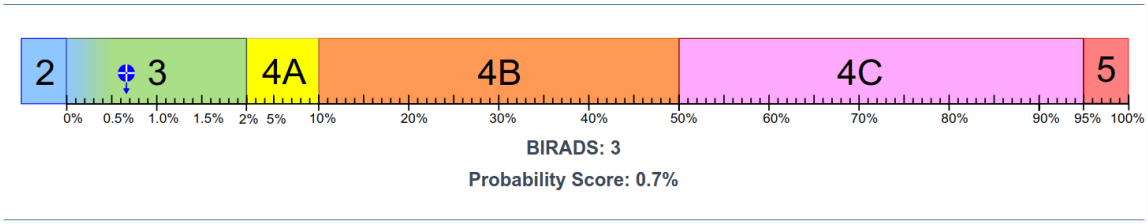
11. GRAPHICAL INTERFACE FOR BI-RADS CATEGORY AND POM RATING

For the purpose of this study, the readers will use a graphical interface to assist them in their assessment of applying their POM to the BI-RADS category for IUS and Imagio (IUS+OA).

The POM scale is linear for all BI-RADS categories; however, the scale for BI-RADS 3 is expanded ten-fold compared to other categories. This design allows for a finer gradation of the malignancy rating for BI-RADS 3 lesions, which in turn provides more data points for the ROC curve and enables a better estimation of the partial AUC. One advantage of a

graphical metaphor is that it provides an aid to the Reader in assigning first a POM score and then a BI-RADS category.

Figure 3: Graphical input of BI-RADS category and POM rating



12. DATA COLLECTION AND QUALITY ASSURANCE

12.1 Data Management

Independent read data for this study is populated in Medidata Rave EDC System via direct data entry during read sessions from readers and read monitor staff. Electronic data collection forms in Rave will be built during study start-up and designed to meet study aims and statistical analysis needs. Electronic case report forms are created based on the study protocol, ACR-CRI scope of work statement, contract, data flows, and data collection needs. The ACR-CRI Data Manager will ensure the eCRFs are tested appropriately, maintain proper version control and all internal and external/sponsor stakeholders approve the content prior to finalization.

Only authorized sponsor employees will have access to study data to the minimum necessary to fulfill a job role or function. User access will require individual log-in credentials, including user IDs and passwords.

12.2 Quality Assurance

12.2.1 Training

Reader training will be provided as described in section 6.2 above.

12.2.2 Quality Control

Monitoring of the readers will be done by ACR in accordance with the procedures outlined under the Reader Study Process Work Flow Section of the protocol and the Independent Review Charter.

12.2.3 Monitoring

Monitoring methods for assuring data quality for each reader, accurate qualification of each reader and reader responsibilities during the read process will be documented in the Independent Review Charter.

13. PARTICIPANT RIGHTS AND CONFIDENTIALITY

13.1 Informed Consent Forms

Not applicable for this study.

13.2 Participant Confidentiality

Reader sets will be de-identified to maintain PIONEER study participant's confidentiality according to the Health Insurance Portability and Accountability Act (HIPAA), any special data security requirements as stipulated by participating readers, and record retention per the sponsor's requirements.

Any data, forms, reports, video recordings, and other records that the participating reader receives from the sponsor will be identified only by a participant identification number (Participant ID, PID) from the PIONEER study to maintain confidentiality. All records will be kept in a locked file cabinet. All computer entry and networking programs will be done using PIDs only. Information will not be released without written permission of the participant, except as necessary for monitoring by the FDA.

13.3 Study Discontinuation

The study may be discontinued at any time by the sponsor for futility or by FDA, or other government agencies as part of their duties to ensure that research participants are protected.

14. PUBLICATION OF RESEARCH FINDINGS

Any presentation, abstract, or manuscript will be made available for review by the sponsor prior to submission.

Study results may be published in support of reimbursement.

Seno will review and approve all manuscripts.

15. REFERENCES

1. *Huang Z, Samuelson F, Tcheuko T, Chen W. Adaptive Design in Multi-reader Multi-case Clinical Trials of Imaging Devices. Statistical Methods in Medical Research, 0(0) 1-20*
2. *Adaptive Designs for Clinical Trials. September 2018.*
3. *Wittes J, Brittain E. The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials. Statistics in Medicine, 9, 65-72, 1990.*
4. *Kieser, M., Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Statistics in Medicine 22:3571-3581.*
5. *Friede T, Kieser M. Sample Size Recalculation in Internal Pilot Study Designs: A Review. Biometrical Journal. 48, 4, 537-555, 2006.*
6. *Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994 Dec; 50 (4): 1029-41.*
7. *Dorfman, DD, Berbaum, KS, & Metz, CE (1992). Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. Investigative Radiology, 27, 723-731.*
8. *Hillis, SL, Obuchowski, NA, Schartz, KM, & Berbaum, KS (2005). A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Statistics in Medicine, 24, 1579-1607 DOI:10.1002/sim.2024.*

9. *Hillis, SL (2007). A comparison of denominator degrees of freedom for multiple observer ROC analysis. Statistics in Medicine, 26:596-619 DOI:10.1002/sim.2532.*

10. *Hillis, SL, Berbaum, KS, & Metz, CE (2008). Recent developments in the Dorfman-Berbaum-Metz procedure for multi-reader ROC study analysis. Academic Radiology, 15, 647-661. DOI:10.1016/j.acra.2007.12.015.*

11. *Hillis, SL, Obuchowski NA, Berbaum, KS (2011). Power estimation for multi-reader ROC methods. Academic Radiology, 18, 129-142. DOI:10.1016/j.acra.2010.09.007.*

12. *Hillis, SL. (2014). A marginal mean ANOVA approach for analyzing multi-reader multi-case radiological imaging data. Statistics in Medicine, 33, 330-360. DOI:10.1002/sim.5926.*

13. *Obuchowski, NA, Rockette, HE (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. Communications in Statistics-Simulation and Computation, 24, 285-308.*

14. *Nofuentes JA, de Dios Luna del Castillo J (2007). Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. Statistics in Medicine 2007 Mar. 26, 4179-4201.*

16. VERSION HISTORY

Version	Date	Description of Changes
1	04Mar19	Original protocol

Version	Date	Description of Changes
2	03Jul 19	• General > Multiple formatting changes and corrections to typographical errors were made throughout the document for clarification > Minor additions of text for clarification throughout the document > "Imagio (OA + IUS + Senogram)" or "Imagio (OA + IUS) Post-Senogram", changed to Imagio (IUS +OA) throughout the document • Page 1 – Clinical Study Protocol Title Page: Indication statement reworded for clarity • Page 2 Tom Stavros title Changed to Chief Medical Officer • Page 6 – List of Abbreviations and Definitions of Terms: added abbreviations and terms used throughout the protocol • Section 4 – Senogram: entire section revised to clarify the Senogram status post the Reader-01 Feasibility Study • Section 5 – Study Design: text revised and or added to clarify reader training • Section 6.2, item 2 Interactive reading, c: deleted "Some malignant masses that could cause false negative Reader-02 Pivotal Study OA reads." • Section 6.2, item 4 Pass/Fail criteria: revised text to clarify pass / fail. • Section 7.2 Exclusion Criteria: deleted "Masses used for development, tuning and training the Senogram. • Section 8.1: revised Figure to reflect Reader-02 Pivotal Study workflow • Section 8.5: deleted "*When possible, the independent QARs will circle the mass of interest for IUS and OA on available mammograms."* • Section 10.1: deleted "This Pivotal Study will further evaluate the primary and secondary endpoints for the benign and malignant subsets with Bi-Rads >0 mammography data. The retrospective powers will be calculated to check if there are sufficient numbers of benign and malignant image sets to reach statistically significant conclusions." • Section 10.7, item 2 Upgrades / Downgrades: deleted entire item • References: deleted reference 8 + 9 and added Chen as 8

3	23 Sep 2019	• General > Minor additions of text throughout document for clarification • Page 5 - List of Abbreviations and Definitions of Terms: added: CB abbreviation and HIPPA • Section 2.1 – Added: A clinically significant gain "of 10%" • Section 2.2 - Revised the order of Secondary Endpoints: NLR, PLR, Partial AUC and deleted the endpoint: Observed and model-adjusted specificities, sensitivities, and paired differences. • Section 5 - Removed "representing current clinical practice"; – Added: (360-10 color balance masses =350); Added language at end of section clarifying the handling of color balance masses included in pivotal study • Section 7.3 - Added clarification "per block to be the same for all readers" • Section 8.1 - Added sentences " The order of masses per block will be the same for each participating reader." "Each reader will be separately proctored by read monitors to ensure that no technical issues arise and that reader scoring is independent." • Section 8.2 - Added sentence "Readers read blocks including IUS and Imagio (IUS+OA) reads in sequential order in accordance to their schedule and read speed." • Section 8.4 - Clarified sentence. "BBRF will generate the randomly selected read list. The read order will be the same for all readers within each block of 120 masses presented." • Section 10 - Added "(360-10 color balance masses=350); Text added on color balance mass exclusion. • Section 10.2 - Text added to add color balance exclusion and added: "Neither futility nor efficacy will be assessed; this is not an adaptive design." • Section 10.4 - Text added to clarify color balance and recommendations for interim analysis added. • Clarification that "No alpha will be spent to perform this interim analysis since this analysis is for sample size re-estimation" • Section 10.7 added - Hierarchal Testing section • Section 10.8 – Added: All analyses will be performed in a GCP-controlled environment using the Hillis software (see references 3-7)"; Clarification of color balance masses; Clarified "a 10% absolute advantage for Imagio vs IUS in specificity for fixed 98% sensitivity is considered to be clinically relevant; this drives the final sample size calculation." "The partial AUC will be computed over the 95-100% sensitivity range. Removed ".005" language. Deleted previous wording on NLR/PLR and added: "NLR and PLR will be analyzed using the logarithmic transformation and the delta method. Calculations will be performed using MATLAB script: DLRATIOS.M. The

Version	Date	Description of Changes
		changes in NLR and PLR correspond to the 10% specificity gain for the primary endpoint.” References: Added Keiser and deleted Chen

4	4Nov2019	• General > Multiple formatting changes and corrections to typographical errors were made throughout the document for clarification • List of Abbreviations and Definitions of Terms: added abbreviations and terms used throughout the protocol • Section 10.2: Reader/Image Set Sample Size: Added the supporting calculations including the software output and added "*in increments of 120 masses per block" to clarify block description and clarified "All original color balance masses will be will be replaced with comparable masses and image sets within the final analysis from the PIONEER Pivotal overage masses.* • Section 10.2: Reader/Image Set Sample Size: Removed "and from the PIONEER Pilot Study as needed." • Section 10.3 Additional Hypothesis Tests added describing hypothesis tests to be done on secondary endpoints: NLP, PLR, Partial ROC/AUC • Section 10.5, 10.5.1 and 10.5.2 Interim Analysis added language to clarify the Plan (10.5.1) and the Methodology (10.5.2) to be used with the interim analysis • Section 10.5.3: Timing: Number of masses added to clarify blocks • Section 10.5.4: Execution: Language clarified " The unblinded biostatistician will only evaluate the overall SD which was the nuisance parameter driving sample size for the fixed 10% difference to be detected. The unblinded biostatistician will not need to take the observed delta into account. Other than the preplanned sample size calculation to determine the remaining blocks required to complete the study, no other information will be revealed to the Sponsor or to any other members of the analysis team. The IA will be executed to the blinded principle." • Section 10.9: Data Analyses: Sentence removed the sentence : "In clinical practice the Probability of Malignancy (POM) Scores (0 100) are standardized such that a BI-RADS 3 (probably benign) shows a likelihood of cancer to be >0% up to 2% inclusive." • Section 10.9: Data Analyses: Added the sentence under 1: "Specificity and sensitivity are calculated using this fixed 2% cutoff. Please be assured that the OR-DBM MRMC 2.51 (Hillis) software does not pick or use a fixed cutoff. In our application, the Hillis software computes AUC, partial AUC, specificity gain at fixed sensitivity, and sensitivity gain at fixed specificity using a full factorial ANOVA incorporating terms for readers, masses, and modalities. The gain in specificity can be computed at a wide range of sensitivities. The 98% sensitivity corresponds to a 2% false negative (FN) rate, but this unrelated to the use of a fixed 2% POM cutoff for clinical decision making." • Section 10.9: Data Analyses: Added the sentence under 2:"The method of Nofuentes (14) will be used; this reflects that the NLR and PLR are both ratios of independent random variables. The DLRATIOS routine computes the standard deviation for the ratios of two independent random variables corresponding to the sensitivity and specificity which

Version	Date	Description of Changes
		are both used to compute NLR and DLR." • References: Added: *Huang Z, Samuelson F, Tcheuko T, Chen W. Adaptive Design in Multi-reader Multi-case Clinical Trials of Imaging Devices. Statistical Methods in Medical Research, 0(0) 1-20* *Adaptive Designs for Clinical Trials. September 2018.* *Wittes J, Brittain E. The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials. Statistics in Medicine, 9, 65-72, 1990.* *Friede T, Kieser M. Sample Size Recalculation in Internal Pilot Study Designs: A Review. Biometrical Journal. 48, 4, 537-555, 2006.* *Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994 Dec; 50 (4): 1029-41.* *Nofuentes JA, de Dios Luna del Castillo J (2007). Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. Statistics in Medicine 2007 Mar. 26, 4179-4201.*

READER STUDY-02

Statistical Analysis Plan

Imagio Pivotal Multi-Reader, Multi-Case Study of Optoacoustic Images versus Imagio Ultrasound to Guide Decision to Biopsy

Philip Lavin PhD, FASA, FRAPS
Roger Aitchison, ScM

For

Seno Medical Instruments
8023 Vantage Dr, Suite 1000
San Antonio, TX 78230

CONFIDENTIALITY STATEMENT

This document contains information that is the property of Seno Medical Instruments Inc. and therefore is provided to you in confidence for review by you, your staff, an applicable Institutional Review Board and regulatory authorities. It is understood that this information will not be disclosed to others without written approval from Seno Medical Instruments Inc., except to the extent necessary to obtain informed consent from those women undergoing evaluation with the Imagio™ system.

Table of Contents

1. ACRONYMS

ACR	American College of Radiology
AE	Adverse Event
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BBRF	Boston Biostatistics Research Foundation
BI-RADS (BR)	Breast Imaging-Reporting and Data System
CB	Color Balancing
CDU	Conventional Diagnostic Ultrasound
CI	Confidence Interval
CSR	Clinical Study Report
EDC	Electronic Data Capture System
FDA	Food and Drug Administration
GCP	Good Clinical Practice
GEE	Generalized Estimating Equations
IA	Interim Analysis
ICF	Informed Consent Form
ICH	International Conference on Harmonization
IRB	Institutional Review Board
ITD	Intention-to Diagnose
IUS	Imagio Internal Ultrasound
MRMC	Multi-reader, Multi-case

NLR	Negative Likelihood Ratio
OA	Opto-acoustics
OR	Obuchowski-Rockette
PD	Protocol Deviation
PLR	Positive Likelihood Ratio
POM	Probability of Malignancy
PMA	Pre Market Approval
ROC	Receiver Operating Characteristic Curve
SAP	Statistical Analysis Plan
SAS	Statistical Analysis System
SG	SenoGram
SSRE	Sample Size Re-Estimation
TPB	Truth Panel Benign

2. STATISTICAL METHODS PLANNED IN THE PROTOCOL AND DETERMINATION OF SAMPLE SIZE

The objective of this analysis plan is to describe the planned full study analysis.

2.1. Statistical and Analytical Plans

2.1.1. Definitions and General Considerations for Data Analysis

1. All calculations, unless otherwise specified, will be performed in a GCP-compliant environment using SAS statistical software (version 9.3 or later), StatXact (version 10 or later), or OR-DBM MRMC (version 2.5 or later).
2. This Statistical Analysis Plan (SAP) is based on Clinical Trial Protocol entitled: Imagio Pivotal Multi-Reader, Multi-Case (MRMC) Study of Optoacoustic Images versus Imagio Ultrasound to Guide Decision to Biopsy.
3. This study is a single arm, controlled, blinded, MRMC study using a sequential design.
4. This pivotal study will be registered at clinicaltrials.gov
5. There will be up to 20 new, independent, prospectively qualified readers to evaluate images using standardized equipment at American College of Radiology

(ACR). These readers will not have previously served as Seno readers for the PIONEER Pivotal or Pilot Study (henceforth referred to as the PIONEER Study), MAESTRO Study, or READER-01 Feasibility Study.

6. All readers will read in accordance with Section 5 of the protocol. Each complete image set is based on a single mass per PIONEER Pivotal Study subject.
7. The image sets for this study will come from the PIONEER Pivotal Study. Image sets are constructed from existing mass-specific images (mammography plus ultrasound and opto-acoustics, alone and in combination). Image sets will be allocated for reader training, testing, and image set reads. Image sets previously used for reader testing or training will remain for testing and training. Image sets used for the READER-02 pivotal study were not previously used for reader testing or training. No READER-02 study image sets will be used for SenoGram (SG) or Color Balance algorithm training. It is estimated that at least 840 image sets will be available for reads and remaining (~900) image sets will be available for SG training.
8. A total of 480 to 840 complete image sets from the original PIONEER Pivotal Study Intention-to Diagnose (ITD) population will be randomly selected from within the PIONEER Pivotal Study.
    - A prospective sampling plan dated February 26, 2019 has been finalized and implemented to select the targeted proportion with and without mammograms for benign masses. [Note: mammograms were present for nearly all malignant masses which reflects real world practice.]
    - Image sets will be organized in blocks of 120 per block with each block containing 75 benign (including three high risk masses defined as atypical ductal hyperplasia, atypical lobular neoplasia, and lobular carcinoma in situ) and 45 malignant masses classified by Conventional Diagnostic Ultrasound (CDU) as BI-RADS (BR) 3 to 5 selected at random in proportion to the original distribution of BI-RADS classifications among subjects in the PIONEER Study.
    - A reserve of up to 60 different overage masses from the PIONEER Pivotal Study will be available for back-up utilization in the event that any masses in the sampling plan cannot be utilized in accordance with the imaging core lab quality assurance process.
    - The prospective sampling plan is solely based on the PIONEER Pivotal Study while reserve cases may also be drawn from the PIONEER Pilot Study if a suitable PIONEER Pivotal Study overage mass match cannot be found.
9. Color balance (CB) masses (n=29) used to tune the Color Balance Algorithm will be replaced for the final analyses.
10. Within the Imagio (IUS +OA) test arm of the study, each study reader will view the SenoGram (SG) following the scoring of the IUS and OA feature scores after viewing the Imagio (IUS+OA) images. The POM and BI-RADS data collected after the SG viewing will be referred to as Imagio (IUS+OA) in the remainder of this document. The initial IUS read without feature scoring will serve as the

internal control, best representing current clinical practice.

11. The primary endpoint is to evaluate the gain in specificity at fixed 98% sensitivity for Imagio (IUS+OA) vs IUS according to a formal hypothesis test to detect a 10% absolute advantage with 80% power. The differences that can be detected with 80% power will also be computed for the secondary endpoint (partial AUC [pAUC]).
12. The sample size calculations conservatively assume that both readers and masses are random effects. The sample size was computed using the following statistical software written by Stephen Hillis and entitled: Multi-Reader Sample Size Program for Diagnostic Studies [http://perception.radiology.uiowa.edu.].
13. For 80% power, a two-sided hypothesis test with a two-sided 5% alpha and 15 readers, a sample size of 854 image sets is needed; the 79% power with 840 image sets is deemed adequate to test the primary hypothesis test; the sample size was calculated using the same intra- and inter-reader standard deviation observed in the READER-01 Feasibility Study.
14. A blinded interim analysis is planned after all readers complete the reading of the first three blocks (360 -10 Color Balance = 350 image sets). The interim analysis will only assess the reader variability associated with the primary endpoint; the sample size may be left at 480 image sets or increased up to 840 image sets based solely on the variability determined from the primary analysis.
15. READER-01 Feasibility Study data will not be combined with the READER-02 Pivotal Study data but may be referenced in the final Pre-market Approval (PMA).
16. Effectiveness analyses (interim and final) will be performed without the CB cases as the primary analysis. The interim analysis will be based on 350 masses without replacement while the final analysis will be based on the sample size to be determined to potentially include all or some of the 29 replacement cases.
17. For completeness, a pre-defined subset of the PIONEER Study tables and listings will be regenerated for the subset of 480 to 840 masses using the READER-02 Pivotal Study data.
18. For completeness, the differences which can be detected with 80% power will be calculated for the primary endpoint (specificity gain at 98% sensitivity) and for the secondary endpoint (pAUC over 95-100% sensitivity).
19. A hierarchical testing strategy will be implemented to extend labeling by testing effectiveness in the following order: gain in specificity for fixed 98% sensitivity, decrease in NLR, gain in PLR, and gain in pAUC over 95% to 100% sensitivity where gain is the difference between the OA/US and IUS values.

2.1.2. Classification of Cancer Status

Study masses will be classified as biopsy-confirmed cancer, biopsy-confirmed benign or Truth Panel benign from the PIONEER Study. Biopsy-confirmed high-risk masses will be included and will be classified as benign for purposes of effectiveness calculations.

2.1.3. Analysis Populations

Safety results for the safety population will be reported for the final masses read as was done for the PIONEER Pivotal Study.

Masses will be selected according to the Mass Sampling Plan. BBRF stratified by diagnosis, original site CDU BI-RADS, and whether the mammogram images were available in selecting the masses at random from the universe of qualified masses analyzed in the PIONEER Pivotal Study. The Sampling Plan depends solely upon clinical study and statistical criteria, including input from the FDA. SenoGram performance was not considered and had no impact on the plan, nor was the SenoGram development group involved. The masses to be sampled excluded masses used in the READER-01 study as well as masses used for training and testing readers. The sampling strategy reflected the overall benign mass (including high risk masses) and malignant mass distribution in the PIONEER Pivotal Study. When subjects had multiple masses, only the first mass was considered. Blocks of 120 masses were constructed to include 75 benign masses (including 3 high risk mases) plus 45 cancer masses according to the stratification noted above.

At the interim timepoint and after study completion, all protocol deviations (PD) will be reviewed. Technical protocol deviations (such as a system error which would not allow the SenoGram to be viewed or a technical issue that will not allow an image to be viewed by a reader) will be handled during a Protocol Deviation Meeting prior to the interim and final analysis; if an image display issue occurs, then that case will be re-presented to a reader once the technical issue is corrected. If the SenoGram would not display to a reader, that case would be excluded from the analysis. Finally, readers not completing all reads (i.e. due to illness or family emergency) for non-study related reasons will be replaced by one of the pre-determined five back-up readers as necessary. Seno will not have access to any data from readers that did not complete all reads. No reader will be excluded as an outlier.

2.1.4. Sample Size Calculations

The sample size will be based on the primary effectiveness endpoint, namely the gain in specificity for fixed 98% sensitivity. This calculation depends upon the number of image sets (one per mass) to be read, the number of readers, as well as the inter-reader variability associated with the underlying models used for analysis.

The test of the primary hypothesis is powered to detect a 10% absolute specificity advantage (superiority) as the alternative hypothesis with no specificity advantage as the null hypothesis. The primary endpoint will be evaluated in the following testable hypotheses:

H_0: $S_{Imagio} = S_{IUS}$

H_1: $S_{Imagio} \neq S_{IUS}$

where S_{Imagio} and S_{IUS} represent specificity, values associated with Imagio (IUS + OA) and IUS.

The sample size calculations conservatively assume that both readers and masses are random effects; an ANOVA model will be used to compute sample size. The sample size was computed using the following statistical software written by Stephen Hillis and entitled: Multi-Reader Sample Size Program for Diagnostic Studies [http://perception.radiology.uiowa.edu.].

For 80% power, a two-sided hypothesis test with a two-sided 5% alpha, and 15 readers, a sample size of 854 image sets is needed to detect a 10% absolute advantage in specificity for Imagio (IUS + OA) vs IUS, assuming a fixed 98% sensitivity and the same intra- and inter-reader standard deviation observed in the Reader-1 Feasibility Study.

The supporting calculations including the software output follow:

```
Results for Feasibility variance estimates

User-supplied parameter or pilot-study values:
            Design : factorial
             Tests : nonequivalence
   Readers and cases : both readers and cases random
        Input values : treat as known
               Alpha : 0.05
        Input Format : OR variance components with error covariances
test*reader var comp : 0.00203538
      Error variance : 0.03571
                Cov1 : 0.01106919
                Cov2 : 0.0061484
                Cov3 : 0.0043063
                  c* : 120

User-supplied desired power, proposed readers & cases values
       Desired Power : 0.8
Proposed max readers : 15
  Proposed max cases : 2000
Proposed min readers : 15
  Proposed min cases : 20

Corresponding OR variance components, covariances, and correlations
test*reader var comp : 0.00203538
      Error variance : 0.03571
                Cov1 : 0.01106919
                Cov2 : 0.0061484
                Cov3 : 0.0043063
                  r1 : 0.309974517
                  r2 : 0.172175861
                  r3 : 0.120590871

Sample Size Results
 effect size :    readers   :     cases    :    power
         0.1 :          15 :         854 :        0.801
```

Initially, 360 masses will be read by all readers for the purposes of performing the interim analysis to assess the sources of variation and to rerun the sample size estimate for the primary effectiveness endpoint. Neither futility nor efficacy will be assessed; this is not an adaptive design.

The sample size may be increased from the minimum of 480 image sets to a maximum of 840 image sets (7 complete blocks) in increments of 120 masses per block. All masses used in the color balance algorithm will be excluded from the final analysis; these masses will be replaced from the PIONEER Pivotal overage masses and the PIONEER Pilot Study as needed.

The maximum sample size to be used for the study will be capped at 840 masses. An overage mass set will be set aside in case quality control issues are found with the images and replacement masses are needed. After excluding the pool of 840 masses and overage masses, approximately

900 masses will remain for independent SenoGram training. No masses used to train the SenoGram, will be in the pool of 840 study masses plus overage. We have removed the 29 masses from the pool of study masses which were mistakenly included from the original color balance algorithm.

Every effort will be made to identify missing reader generated data in real time. Each block of 120 masses will include 75 benign (72 benign + 3 high risk) and 45 malignant masses to represent the PIONEER study.

With the exception of masses excluded due to protocol deviations, all complete reads will be included in all analyses, consistent with an "Intention-to-Diagnose" approach.

2.1.5. Disposition of Masses

A total of 840 masses with 120 masses per block have been selected from the PIONEER Pivotal Study ITD population. The additional 60 overage masses allowed for replacement of any mass that might prove unreadable. Replacement of any masses will be documented in the Clinical Study Report (CSR).

2.1.6. Demographic and Other Baseline Characteristics

Clinical presentation and medical history (limited to the specific masses used from the PIONEER database), including the referral indication, palpability of mass, breast density, age, menopausal status, presence of breast implants, number of masses (in PIONEER), and mammographic intent will be summarized using descriptive statistics.

Summary tables will display demographic and baseline characteristics by diagnostic status (benign+TPB, cancer); TBP were the masses without biopsy which were declared to be Truth Panel Benign; high risk masses will be included with the benign masses.

2.1.7. Medical History and Concomitant Medications

Detailed medical history (as presented to the readers) will be listed. Concomitant medications are irrelevant to this study.

2.1.8. Effectiveness Methodology

The following approaches (Hillis OR-DBM MRMC 2.51) were used to treat readers as correlated and as random effects in analyses. [redacted]

[redacted]

Analysis 1 treats both as random, and thus results generalize to both the population of readers and cases – this is the situation for which the DBM and Obuchowski-Rockette (OR) procedures were originally designed. However, within the OR/DBM analysis framework one can also analyze the

data treating only cases as random (Analysis 2) or readers as random (Analysis 3), although we note that these analyses are not unique to OR/DBM. Analyses 2 and 3 are viewed as supportive analyses while Analysis 1 is viewed as the primary analysis approach.

For all three analyses, the null hypothesis of equal treatments is tested in part (a), diagnostic test difference 95% confidence intervals are given in part (b), and individual diagnostic test two-sided 95% confidence intervals are given in part (c) of the output. For the DBM procedure, parts (a) and (b) are based on the diagnostic test x reader x mass ANOVA of jackknife pseudo-values while part (c) is based on the reader x case ANOVA of jackknife pseudo-values for the specified diagnostic test. For the OR procedure, parts (a) and (b) are based on the diagnostic test x reader ANOVA of reader-performance outcomes (e.g., AUCs) while part (c) is based on the reader ANOVA of reader-performance outcomes for the specified diagnostic test. [redacted]

Different denominator "error terms" (these are the denominators used for the F statistics) are used as indicated for parts (a), (b), and (c) according to whether reader and case are treated as fixed or random factors, and their formulas are provided in the output. Note that the diagnostic test confidence intervals in part (c) are based only on the data for the specified diagnostic test, rather than the pooled data. Diagnostic test difference 95% confidence intervals for each reader are presented in part (d) of Analysis 2: for DBM each interval is based on the diagnostic test x mass ANOVA table for the specified reader; for OR each interval is based on the OR variance and Cov1 estimates computed separately for each reader.

Analysis 1

Our planned analysis (Analysis 1) treats both readers and masses as random samples. Results apply to the reader and mass populations — this is the situation for which DBM and OR were initially created.

Analysis 2

Analysis 2 treats only masses as a random sample. Results apply to the population of masses but only for the readers used in the study. For this analysis, inferences are based on the estimated error covariance matrix, treating readers as fixed. These two methods will give almost the same results for typical studies where the total number of cases is at least moderate (≥ 50). Diagnostic test differences and two-sided 95% confidence intervals are presented for each reader in part (d).

Analysis 3

Analysis 3 treats only readers as a random sample. Results apply to the population of readers but only for the masses used in this study. These results are based on a conventional diagnostic test x reader ANOVA for the AUCs (or other measures), where reader is a random factor and diagnostic test is a fixed factor which is included for completeness. This is the same as a repeated measures ANOVA where diagnostic test is the repeated measures factor, i.e., readers provide an outcome (e.g., AUC) for each diagnostic test. For two tests, this analysis is equivalent to a paired t test performed on the reader-specific AUC estimates. DBM and OR give the same results here, and it

does not matter with error covariance method is used with OR because the covariance does not enter into the F statistic.

See the respective equations for variance components and degrees of freedom (DF) in Appendix 1.

2.1.9. Effectiveness Analysis

There is a formal test of hypotheses for the primary endpoint. Other p-values that are generated for other endpoints will be regarded as descriptive statistics. Imagio (IUS+OA) represents the reader input after viewing the Imagio (IUS+OA) images and SenoGram results. IUS represents the reader input prior to any IUS feature scoring, but after reader training on Imagio (IUS+OA).

The following will be used to assess the pivotal study.

Table 1: Effectiveness Analysis for Study Endpoints

Endpoints by Cohort	Analysis/Software
Primary	
Gain in specificity Imagio (IUS+OA) vs. IUS at fixed 98% sensitivity interpolated from the ROC AUC curves using methods of Obuchowski-Rockette (OR). Effect estimates (two-sided 95% CIs) will be derived using OR-DBM MRMC methods. The interim analysis to re-assess pivotal study sample size will be determined using OR-DBM MRMC sample size program.	OR-DBM MRMC Sample Size: OR-DBM-MRMC Power program, nQuery v7, SAS (PROC POWER), and/or PASS v15.0.5.

Secondary	
Partial AUC: Difference (Imagio (IUS+OA) vs. IUS) in partial Receiver Operator Characteristic (ROC) Area Under the Curve (AUC) for 95-100% sensitivities using the OR methods.	OR-DBM
Specificity and Sensitivity: Specificity is defined using the percent with a negative result (POM <2%) among all benign+TPB masses (to include all high-risk masses). Sensitivity is defined using the percent with a positive result (POM >2%) among all malignant masses. Observed and model-adjusted specificity and sensitivity will be reported for Imagio (IUS+OA) vs. IUS using a 2% Probability of Malignancy (POM) cutoff. These analyses will be repeated for POM cutoffs of 1%, 3%, 4%, 5%, 6%, and 7%. High risk masses will be included in the calculations for specificity.	SAS PROC GEE and MIXED treating masses as independent observations and readers as correlated and then readers as independent and masses as correlated. Two-sided 95% confidence intervals (CIs) for sensitivity and specificity will be computed for IUS, for Imagio (IUS+OA), and for the pairwise difference. Confidence intervals will be constructed treating masses as independent observations with readers as correlated and then readers as independent with masses correlated.
Negative Likelihood Ratios (NLR) and Positive Likelihood Ratio (PLR) for IUS and Imagio (IUS+OA), NLR defined as ((1- observed sensitivity)/observed specificity), and relative NLR and PLR for paired designs.	SAS: Univariate 95% CI for PLR and NLR with variances fit using the logarithmic transformation and the delta method, reference PIONEER Table 14.2.3.2. MATLAB script: [redacted]

SenoGram performance using the following metrics from the study endpoints: a) Sensitivity and specificity for SenoGram classification based on predicted probability and the SenoGram threshold, with subgroup analysis for masses with and without mammogram BI-RADS data. b) Specificity at fixed sensitivity (98%) for SenoGram classification as in (a). c) Partial AUC (over 95% to 100% sensitivity, inclusive) for SenoGram classification as in (a).	For each metric, a summary table for each reader (where appropriate) and for all readers to contain the following: • Measured value using READER-02 feature data. • Estimated value from SenoGram cross-validation with PIONEER feature data. • Estimated value from SenoGram predictions for masses with PIONEER data.
Exploratory	
OA Feature Score Distributions: • Benign vs. malignant masses • Benign masses only – mammogram vs no mammogram POM Score Distributions: • Benign vs. malignant masses • Benign masses only – mammogram vs no mammogram Inter-reader variability assessments (POM and feature scoring): • Mammogram • IUS • OA.	Summary Statistics by reader. Average over readers and then summarize means for overall. Use a Wilcoxon Rank Sum test. Reference PIONEER T14.2.10.1 [t_14_02_1x_feat-10-_01] Reference PIONEER Table 14.2.9.1 [t_14_02_09_pom-01] Statistics will be calculated to assess inter-reader variabilities.

2.1.10. Hypotheses to be Tested

The test of the primary hypothesis is powered to detect a 10% absolute specificity advantage (superiority) as the alternative hypothesis with no specificity advantage as the null hypothesis. The primary endpoint will be evaluated in the following testable hypotheses:

H_0: $S_{Imagio} = S_{IUS}$

H_1: $S_{Imagio} \neq S_{IUS}$ representing a 10% absolute increase

where S_{Imagio} and S_{IUS} represent specificity, values associated with Imagio (IUS + OA) and IUS.

The following hypothesis tests will be applied using the observed estimates for the following secondary endpoints:

- NLR:
  - H_0: $NLR_{IUS} = NLR_{Imagio}$ vs
  - H_A: $NLR_{IUS} \neq NLR_{Imagio}$ representing a reduction
- PLR:
  - H_0: $PLR_{IUS} = PLR_{Imagio}$ vs
  - H_A: $PLR_{IUS} \neq PLR_{Imagio}$ representing an increase
- Partial ROC AUC:
  - H_0: $pAUC_{IUS} = pAUC_{Imagio}$ vs
  - H_A: $pAUC_{IUS} \neq pAUC_{Imagio}$ representing an increase

Hierarchical testing will be applied to seek labeling claims if warranted by the pre-defined testing procedure (see Section 2.1.11.4).

2.1.11. Data Analysis Methodology

Data Analyses

As for the PIONEER Pivotal Study, analyses will be based on independent readers for POM, BI-RADS and feature scores (IUS, Imagio (IUS+OA)) individually and overall. There is a formal hypothesis test for the primary endpoint and the three secondary endpoints. Hierarchical testing will be in place to control the Type I error. Truth (determination of malignant biopsy, benign biopsy, HR or TPB) will be taken from the PIONEER Pivotal Study. All CB masses (n=29) will be excluded from all analyses; these masses will be replaced in order to preserve final analysis power.

The specificity advantage of Imagio (IUS+OA) vs. IUS will be compared at fixed 98% sensitivity. The software package OR-DBM MRMC [Version 2.5 or later] will be used for this analysis of OA gain as interpolated from the ROC curves.

Specificity and sensitivity will be calculated by reader and overall for IUS, Imagio (IUS+OA), and the pairwise difference [PROC GENMOD, see Table 1]. Specificity and sensitivity will be calculated using the 2% POM cutoff for IUS and for IUS+OA; and then for 1%, 3%, 4%, 5%, 6%, and 7% cutoffs to assess robustness. In clinical practice, the POM Scores (0-100) are standardized such that a BI-RADS 3 (probably benign) shows a likelihood of cancer to be > 0% up to 2% inclusive.

Specificity and sensitivity are calculated using this fixed 2% cutoff. Please be assured that the OR-DBM MRMC 2.51 (Hillis) software does not pick or use a fixed cutoff. [redacted] The gain in specificity can be computed at a wide range of sensitivities. The 98% sensitivity corresponds to a 2% false negative (FN) rate, but this unrelated to the use of a fixed 2% POM cutoff for clinical decision making.

NLR and PLR will be analyzed using the logarithmic transformation and the delta method. Calculations will be performed using MATLAB script: [redacted]

[redacted]

Last, pAUC for fixed 95% to 100% sensitivity will also be analyzed using OR-BDM MRMC.

The sample size for the READER-02 pivotal study was calculated using method implemented in the MRMC Sample Size Program 1.0 for Diagnostic Studies, by Hillis, Obuchowski, and Birnbaum. These sample size results may be confirmed by other packages such as nQuery, PASS, and SAS [PROC POWER].

POM summary statistics for IUS, Imagio (IUS+OA), and the pairwise difference will be displayed by diagnostic category and compared using a Wilcoxon Rank Sum test. POM scores for benign masses will be displayed for masses with and without mammograms.

OA feature scores including internal total, external total, and overall total scores will be summarized to describe the distribution of individual feature scores by diagnosis, with a Wilcoxon statistical test for the difference in the distributions between malignant and benign. Feature scores for benign masses will be displayed for masses with and without mammograms.

2.1.11.1. Adjustments for Covariates

No adjustments for covariates are planned.

The primary and secondary endpoints will be rerun for the subgroups with and without mammograms.

2.1.11.2. Handling of Dropouts or Missing Data

Missing reader data will be reported (see protocol deviations). No imputation is planned for any excluded study data. The analysis methods can accommodate missing reader data.

2.1.11.3. Interim Analyses and Data Monitoring

Plan

The interim analysis (IA) is to re-estimate sample size to avoid running an underpowered study. The IA will be performed by an unblinded statistician familiar with the study data. The Sponsor and all other parties will remain blinded to the results. The only communication to the Sponsor will be the final sample size recommendation. The sample size may be increased up to 840 masses from the initial 480 masses in order to achieve 80% power for the primary endpoint. This recommendation will be based solely on variance parameters re-estimated in the IA (but not the effect size). Thus, these IA results fall into the category of non-comparative results. Thus, the IA has negligible effect on the Type 1 error. The IA will be based on the total variance estimate (inclusive of both reader and imaging modality variations) so the IA remains 'non-comparative' since the interim treatment effect was not used to compute the final sample size.

Methodology

The sole IA goal is to compute the final sample size based on the primary endpoint. The sample size re-estimation (SSRE) will be performed without any adaptation. SSRE will be conducted to detect the 10% absolute advantage in specificity for fixed 98% sensitivity, given that this is an established criterion agreed upon with FDA. The standard deviation (from reader as a random effect) is not known. There will be no preplanned or performed

assessments for futility, stopping for early efficacy, dropping readers, altering the 10% effect hypothesis, or any other adaptation. Therefore, this is classic SSRE.

Huang and Chen (1) reference this application as distinct from his discussion of adaptive designs. Our IA is based entirely on non-comparative data, that is, no information about treatment effects were incorporated into the decision process. For this SSRE situation (a non-comparative interim), it can be shown that performing analyses at the conventional one-sided 0.025 significance level has a negligible effect on the Type 1 error probability (FDA Guidance on Adaptive Trials (2), Wittes (3), Kieser and Friede (4), Friede and Kieser (5), among others). Wittes advises that the alpha spend is small when the observed variance is less than the assumed variance (as per the interim analysis) without accounting for imaging modality. Huang and Chen advise that the alpha spend is small when the conditional power exceeds 60% which is the case when a 10% delta is assumed; any delta effect would further reduce the reader variance. They specifically state that "for conditional power within a range of 0.37 and 0.80, the sample size can be increased while leaving the critical test value unchanged"; this is the fundamental premise of the promising zone. Thus, the alpha spend for conducting the planned interim analysis is small.

From a testing perspective, the conclusions should hold in recognition that the p-value is a Z-score whether derived from a one-sample or a two-sample test independent of the model used (Huang or Hillis). All alpha adjustment depends on a linear combination of Z-scores (6) for respective Z-scores prior to and after the pre-planned interim analysis. Thus, first principles all lead to the conclusion that the alpha spend is small enough to ignore. Our estimate is that the alpha spend is no more than 0.01% under the given conditions. The minimal alpha spend derives from the abundant two-sample literature to apply to our paired sample setting with no published literature.

Timing

A blinded interim analysis is planned for this study after all readers complete the reads for the first three blocks (360 - 10 CB masses = 350 complete image sets). The CB masses (n=10) will be excluded from the interim analyses.

The interim analysis will solely re-estimate sample size based on the primary endpoint; it will not be used to modify the alternative hypothesis. Neither futility nor efficacy will be assessed. The following recommendations will be issued to Seno by the unblinded statistician at the interim timepoint:

1. Sample Size Calculation:
   - ___One additional block (480 masses) is required OR
   - ___Two additional blocks (600 masses) are required OR
   - ___Three additional blocks (720 masses) are required OR
   - ___Four additional blocks (840 masses) are required OR
   - ___Four additional blocks (840 masses) are not sufficient but reads should continue because there is a reasonable chance to reach statistical significance based on the conditional power assuming the alternative hypothesis holds.
   - ___Four additional blocks (840 masses) are not sufficient, reads should not continue because there is not a reasonable chance to reach statistical significance

based on the conditional power assuming the alternative hypothesis holds

Execution

The unblinded biostatistician will only evaluate the overall standard deviation which was the nuisance parameter driving sample size for the fixed 10% difference to be detected. The unblinded biostatistician will not need to take the observed delta into account. Other than the preplanned sample size calculation to determine the remaining blocks required to complete the study, no other information will be revealed to the Sponsor or to any other members of the analysis team. The IA will be executed to the blinded principle.

2.1.11.4. Multiple Comparisons / Multiplicity

To extend labeling claims beyond the primary endpoint, a hierarchical testing strategy will be used to control Type 1 error associated with testing the primary endpoint and three secondary endpoints in the following pre-defined order until statistical significance is no longer reached in support of extended labeling:

- Increase in specificity for fixed 98% sensitivity.
- Decrease in NLR
- Increase in PLR
- Increase in pAUC over 95% to 100% sensitivity

No other endpoints are to be included as labeling claims.

2.1.11.5. Examination of Subgroups

Study endpoints for benign mases with and without mammograms will be presented.

2.1.12 Evaluating the SenoGram

SenoGram evaluation is divided into 1) evaluation of reader utilization, and 2) correctness of the SenoGram prediction. Evaluation of reader utilization includes the following analysis:

- N (%) masses with reader reporting that SenoGram increased their confidence in their assigned BI-RADS category.
- N (%) masses with reader reporting that SenoGram decreased their confidence in their assigned BI-RADS category.

Evaluation of correctness of SenoGram predictions will be based on the analysis listed in Table 1.

2.1.13. Safety Evaluation

2.1.13.1. Extent of Exposure

OA Exposure for subjects in this study will be reported as was done in the. PIONEER CSR.

2.1.13.2. Adverse Events

Adverse Events for subjects in this study will be reported as was done in the. PIONEER CSR.

2.1.13.3. Clinical Laboratory Evaluation

No clinical laboratory evaluations are made in this study.

No vital signs, physical findings, and other safety observations are made in this study.

2.2. Hypotheses to be Tested

See protocol Section 10.2 for the formal hypotheses to be tested.

2.3. Determination of Sample Size

See protocol Section 10.2 for the sample size justification.

3. CHANGES IN THE CONDUCT OF THE STUDY OR PLANNED ANALYSES

None.

4. REFERENCES:

1. Huang Z, Samuelson F, Tcheuko T, Chen W. Adaptive Design in Multi-reader Multi-case Clinical Trials of Imaging Devices. Statistical Methods in Medical Research, 0(0) 1-20
2. Adaptive Designs for Clinical Trials. September 2018.
3. Wittes J, Brittain E. The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials. Statistics in Medicine, 9, 65-72, 1990.
4. Kieser, M., Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Statistics in Medicine 22:3571-3581.
5. Friede T, Kieser M. Sample Size Recalculation in Internal Pilot Study Designs: A Review. Biometrical Journal. 48, 4, 537-555, 2006.
6. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994 Dec; 50 (4): 1029-41.
7. Dorfman DD, Berbaum KS, Metz CE (1992). Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. Investigative Radiology, 27, 723-731.
8. Obuchowski NA, Rockette HE (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. Communications in Statistics-Simulation and Computation, 24, 285-308.
9. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. (2005). A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Statistics in Medicine, 24, 1579-1607 DOI:10.1002/sim.2024.
10. Hillis SL. (2007). A comparison of denominator degrees of freedom for multiple observer ROC analysis. Statistics in Medicine, 26:596-619 DOI:10.1002/sim.2532.
11. Hillis SL, Berbaum KS, and Metz CE. (2008). Recent developments in the Dorfman-Berbaum-Metz procedure for multi-reader ROC study analysis.

Academic Radiology, 15, 647-661. DOI:10.1016/j.acra.2007.12.015B

12. Hillis SL, Obuchowski NA, Berbaum KS. (2011) Power estimation for multi-reader ROC methods: An updated and unified approach. Academic Radiology; 18:129-142. doi: DOI: 10.1016/j.acra.2010.09.007.
13. Multi-Reader Sample Size Program for Diagnostic Studies, available at http://perception.radiology.uiowa.edu/Software/PowerandSampleSizeEstimation/tabid/25 2/Default.aspx.M
14. Nofuentes JA and de Dios Luna del Castillo J. Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. Statistics in Medicine; Volume 26, pages 4179-4201 March 2007.

[redacted]

SUMMARY OF CHANGES:

Version	Date	Description of Changes
1.0	02May19	Original SAP

V 2.0	01Jul 19	START V2.0 • Signature Page: Roger Aitchison title change to Sr. Biostatistician, Tom Stavros Title Change to Chief Medical Officer • Table of Contents: Administrative Changes including addition of Appendix 1 and 2 • Acronyms: Administrative additions • Section 2.1.1: Number 2: Removal of "dated March 4, 2019- Version 1.0, subject to modification following a forthcoming FDA submission." • Section 2.1.1: Removal of number 4, "This document number is PROT-00000120." • Section 2.1.1: Addition of new number 7 "The image sets for this study will come from the PIONEER Pivotal Study. Image sets are constructed from existing mass-specific images (mammography plus ultrasound and opto-acoustics, alone and in combination). Image sets will be allocated for training, testing, and image set reads. Image sets previously used for testing or training will remain for testing and training. In contrast, image sets used for the READER-02 study must not have been previously used for testing or training. No READER-02 study image sets will be used for SenoGram (SG) algorithm training; it is estimated that at least 840 image sets will be available for reads and remaining (~900) image sets will be available for SG training." • Section 2.1.1: Number 8: Addition to first bullet point: "which reflects real world clinical practice." • Section 2.1.1: Number 8: Addition to third bullet: Addition of "in accordance with the imaging core lab quality assurance process." • Section 2.1.1: Addition of new number 9: "Within the Imagio (IUS+OA) test arm of the study, each study reader will view the SenoGram (SG) following the scoring of the IUS and OA feature scores after viewing the IUS+OA images. The POM and BI-RADS data collected after the SG viewing will be referred to as IUS+OA in the remainder of this document." • Section 2.1.1: Addition of language to old number 8 to new Number 10: "The primary endpoint is to evaluate the gain in specificity at fixed 98% sensitivity for IUS+OA post vs IUS according to a formal hypothesis test to detect a 10% absolute advantage with 80% power. The differences that can be detected with 80% power will also be computed for each of the secondary endpoints (partial ROC AUC, downgrades, and upgrades); the calculations will not be performed for either sensitivity and specificity to avoid misinterpretation relative to the primary endpoint (which already evaluates the specificity advantage for fixed sensitivity)." • Section 2.1.1: Number 13: Cui language removed related to alpha spend

		and the following language added: "The method of Chen will be considered to reassess the sample size; under no circumstances will the alternative hypothesis be modified as a result of this analysis." • Section 2.1.1: Number 16 added: "For completeness, the differences which can be detected with 80% power will be calculated for each primary and secondary endpoint with the exception of sensitivity and specificity. Clinically meaningful differences follow a. partial ROC AUC: a 0.005 difference (for 95-100% sensitivity) b. downgrades: a 10% advantage, and c. upgrades: rule out a 5% disadvantage." • Section 2.1.3 Wording added " BBRF stratified by diagnosis, original site CDU BI-RADS, and whether the mammogram images were available in selecting the masses at random from the universe of qualified masses analyzed in the PIONEER Pivotal Study. The Sampling Plan depends solely upon clinical study and statistical criteria, including input from the FDA. SenoGram performance was not considered and had no impact on the plan, nor was the SenoGram development group involved. The masses to be sampled excluded masses used in the READER-01 study as well as masses used for training and testing readers. The sampling strategy reflected the overall benign mass (including high risk masses) and malignant mass distribution in the PIONEER Pivotal Study. When subjects had multiple masses, only the only first mass was considered. Blocks of 120 were constructed to include 75 benign masses (including 3 high risk mases) plus 45 cancer masses according to the stratification noted above." • Section 2.1.3 Wording Added re: protocol deviations "Technical protocol deviations such as a system error which would not allow the SenoGram to be viewed or a technical issue that will not allow an image to be viewed by a reader will be handled during a Protocol Deviation Meeting prior to the interim and final analysis. If an image display issue occurs, then that case will be re-presented to a reader once the technical issue is corrected. If the SenoGram would not display to a reader, that case would be excluded from the analysis. Finally, readers not completing all reads (i.e. due to illness or family emergency) for non-study related reasons will be replaced by one of the pre-determined five back-up readers as necessary. Seno will not have access to any data from readers that did not complete all reads. No reader will be excluded as an outlier." • Section 2.1.7 was rewritten to state: "The following approaches (Hillis OR-DBM MRMC 2.51) were used to treat readers as correlated and as random effects in analyses. See Appendix 1 for OR User Manual Section 9 for the underlying models and corresponding variance component and degree of freedom equations including notations. Analyses for three study designs are presented in Appendix 2 from the READER-01 analyses for the following scenarios with primary focus on the standard modality-by-reader-mass factorial study design, where each reader reads

		all masses for both diagnostic tests. See Attachment 1 for discussion on using GEE to estimate specificity in SAS using READER-01 data. Analysis 1 treats both as random, and thus results generalize to both the population of readers and cases – this is the situation for which the DBM and Obuchowski-Rockette (OR) procedures were originally designed. However, within the OR/DBM analysis framework one can also analyze the data treating only cases as random (Analysis 2) or readers as random (Analysis 3), although we note that these analyses are not unique to OR/DBM. For all three analyses, the null hypothesis of equal treatments is tested in part (a), diagnostic test difference 95% confidence intervals are given in part (b), and individual diagnostic test two-sided 95% confidence intervals are given in part (c) of the output. For the DBM procedure, parts (a) and (b) are based on the diagnostic test x reader x mass ANOVA of jackknife pseudo-values while part (c) is based on the reader x case ANOVA of jackknife pseudo-values for the specified diagnostic test. For the OR procedure, parts (a) and (b) are based on the diagnostic test x reader ANOVA of reader-performance outcomes (e.g., AUCs) while part (c) is based on the reader ANOVA of reader-performance outcomes for the specified diagnostic test. These ANOVA tables are displayed before the analyses. Different denominator "error terms" (these are the denominators used for the F statistics) are used as indicated for parts (a), (b), and (c) according to whether reader and case are treated as fixed or random factors, and their formulas are provided in the output. Note that the diagnostic test confidence intervals in part (c) are based only on the data for the specified diagnostic test, rather than the pooled data. Diagnostic test difference 95% confidence intervals for each reader are presented in part (d) of Analysis 2: for DBM each interval is based on the diagnostic test x mass ANOVA table for the specified reader; for OR each interval is based on the OR variance and Cov1 estimates computed separately for each reader. • Analysis 1: Analysis 1 treats both readers and masses as random samples. Results apply to the reader and mass populations — this is the situation for which DBM and OR were initially created. • Analysis 2: Analysis 2 treats only masses as a random sample. Results apply to the population of masses but only for the readers used in the study. For this analysis, inferences are based on the estimated error covariance matrix, treating readers as fixed. These two methods will give almost the same results for typical studies where the total number of cases is at least moderate (≥ 50). Diagnostic test differences and two-sided 95% confidence intervals are presented for each reader in part (d). • Analysis 3: Analysis 3 treats only readers as a random sample. Results apply to the population of readers but only for the masses used in this study. These results are based on a conventional diagnostic test x reader ANOVA for the AUCs (or other measures), where reader is a random factor and diagnostic test is a fixed factor which is included for

		completeness. This is the same as a repeated measures ANOVA where diagnostic test is the repeated measures factor, i.e., readers provide an outcome (e.g., AUC) for each diagnostic test. For two tests, this analysis is equivalent to a paired t test performed on the reader-specific AUC estimates. DBM and OR give the same results here, and it does not matter with error covariance method is used with OR because the covariance does not enter into the F statistic. See the respective equations for variance components and degrees of freedom (DF) in Appendix 1: • Section 2.1.18 Table 1 from Section 2.1.17 SAP v1 moved to Section 2.1.8 and language added under Primary Gain in Specificity endpoint re: Obuchowski and Rockette (OR). "Effect estimates (two-sided 95% CIs) will be derived using OR-DBM MRMC methods." • Section 2.1.8 Secondary Endpoint for Sensitivity and Specificity clarification language added "Specificity is defined using the percent with a negative result (POM <2%) among all benign+TPB masses to include all high-risk masses. Sensitivity is defined using the percent with a positive result (POM >2%) among all malignant masses. High risk masses will be included in the calculations for specificity." • Section 2.1.8 Secondary Endpoint: Downgrade/Upgrade Endpoints removed • Section 2.1.8 Exploratory Endpoint Added: "Inter-reader variability assessments (POM, BI-RADS, OA feature scoring): • Mammogram • IUS • OA. Kappas will be calculated to assess inter-reader variabilities." • Section 2.1.9 First paragraph on primary endpoint alpha spend the "(see Chen)" was added. • Section 2.1.9 Downgrade Upgrade language was removed • Section 2.1.9 "and considering the Chen R program" was added to third to last paragraph on sample size. • Section 2.1.9.4: Multiple Comparisons/Multiplicity: The following language was added: "There is no alpha-splitting or hierarchical testing plan. Additional endpoints are not to be included as labeling claims." • Section 3: The following wording was added "If the Chen R program gives the same answers as the Hillis software, then we will consider using the software at the time of the interim analysis." • Section References: The following reference added "Chen W, Huang Z, Samuelson F, Tcheuko L (2019). Adaptive sample size reestimation in MRMC studies, SPIE 10952, Medical Imaging: Image Perception, Observer Performance, and Technology Assessment, 109520G • Section References: The reference to Cui was removed. • Appendix 1 added: **OR USER MANUAL SECTION 9**

Version	Date	Description of Changes
		• Appendix 2 added: **SAMPLE OUTPUT FOR SPECIFICITY ADVANTAGE** • Various administrative changes END OF V2.0

V 3.0	13Aug2019	START of V3.0 • Acronyms: BBRF was added • Section 2.1.1: Number 10: Sentence was modified to remove downgrade and upgrade language and ROC. It reads …for the secondary endpoint "(partial AUC [pAUC]); the calculations will not be performed for either sensitivity or specificity to avoid misinterpretation relative to the primary endpoint (which already evaluates the specificity advantage for fixed sensitivity • Section 2.1.1: Number 14: "Reader-01" was added for clarification • Section 2.1.1: Number 15: Was revised to read "For completeness, a pre-defined subset of the PIONEER Study tables and listings will be regenerated for the subset of 480 to 840 masses using the READER-02 pivotal study data." • Section 2.1.1: Number 16: Removed letters a,b,c related to partial ROC:AUC, downgrades and upgrades. Is now worded as: "For completeness, differences which can be detected with 80% power will be calculated for the primary endpoint (specificity gain at 98% sensitivity) and for the secondary endpoint (pAUC for the sensitivity range of 95-100%)." • Section 2.1.5: "Last sentence was revised to "Summary tables will display demographic and baseline characteristics by diagnostic status (benign+TPB, cancer); TPB were the masses without biopsy which were declared to be Truth Panel Benign; high risk masses will be included with the benign masses." • Section 2.1.7: Analysis 1 section: "Our planned analysis (Analysis 1) treats both readers…" • Section 2.1.8: Table 1: Secondary: Partial AUC clarified "Partial AUC: Difference (Imagio (IUS+OA) vs. IUS) in partial Receiver Operator Characteristic (ROC) Area Under the Curve (AUC) for 95-100% sensitivities using the OR methods. Differences will be absolute differences (|pAUC(IUS) - pAUC(OA+IUS) with no normalization." • Section 2.1.8 Table 1: Secondary: Analysis/Software: Partial AUC: MRMC & methods as per McClish removed. • Section 2.1.8 Table 1: Secondary: Analysis/Software: Sensitivity and Specificity methods clarified to be "SAS PROC GEE and MIXED treating masses as independent observations and readers as correlated and then readers as independent and masses as correlated." • Section 2.1.8 Table 1: Secondary: Analysis Software: Sentence on Confidence Intervals modified to read "Confidence intervals will be constructed treating masses as independent observations with readers as correlated and then readers as independent with masses correlated.

Version	Date	Description of Changes
		• Section 2.1.8: Table 1: Secondary: SenoGram Performance: Item c removed McClish • Section 2.1.8: Table 1: Exploratory: Analysis/Software: OA Feature Score Analysis and Software modified to: " Use a Wilcoxon Rank Sum test" • Section 2.1.8: Table 1: Exploratory: BI-RADS removed from Inter-Reader variability assessments • Section 2.1.8: Table 1: Exploratory: Analysis/Software: Inter-reader variability modified to "Intraclass correlation coefficient (ICC) statistics will be calculated to assess inter-reader variabilities." • Section 2.1.9: Analsyes: "(see Chen)" added to first paragraph • Section References: Added Number 10 reference "Statistical Assessment Methodology for Diagnostics and Biomarkers: https://www.fda.gov/medicaldevices/ cdrh-research-Programs/statistical-assessment-methodology-diagnostics-and-biomarkers https://github.com/DIDSR/iMRMC/releases" • Various administrative changes END OF V3

V 4.0	23Sep2019	START of V4.0 • Section 2.1.1: Clarification to SAS software: "version 9.3" • Section 2.1.1: Number 7: Revised to read: "Image sets previously used for reader testing or training will remain for testing and training. Image sets used for the READER-02 pivotal study were not previously used for reader testing or training. No READER-02 study image sets will be used for SenoGram (SG) or Color Balance algorithm training. It is estimated that at least 840 image sets will be available for reads and remaining (~900) image sets will be available for SG training." • Section 2.1.1: Number 8: High Risk language removed "DCIS, lymphoma, or phyllodes masses" and replaced with "atypical ductal hyperplasia, atypical lobular neoplasia, and lobular carcinoma in situ" • Section 2.1.1: Added new Number 9 " Color balance (CB) masses (n=29) used to tune the Color Balance Algorithm will be replaced for the final analyses." • Section 2.1.1: Number 11: Wording removed " the calculations will not be performed for either sensitivity or specificity to avoid misinterpretation relative to the primary endpoint (which already evaluates the specificity advantage for fixed sensitivity)." • Section 2.1.1: Number 14: Modified to read "A blinded interim analysis is planned after all readers complete the reading of the first three blocks (360 -10 Color Balance = 350 image sets). The interim analysis will only assess the reader variability associated with the primary endpoint; the sample size may be left at 480 image sets or increased up to 840 image sets based solely on the variability determined from the primary analysis." • Section 2.1.1: Number 16 was added "Effectiveness analyses (interim and final) will be performed without the CB cases as the primary analysis. The interim analysis will be based on 350 masses without replacement while the final analysis will be based on the sample size to be determined to potentially include all or some of the 29 replacement cases." • Section 2.1.1: Number 19 was added: "A hierarchical testing strategy will be implemented to extend labeling by testing effectiveness in the following order: gain in specificity for fixed 98% sensitivity, NLR, PLR, and pAUC over 95% to 100% sensitivity." • Section 2.1.2: Added last sentence: "Biopsy-confirmed high-risk masses will be included and will be classified as benign for purposes of effectiveness calculations." • Section 2.1.7: Last sentence of third paragraph added for clarification "Analyses 2 and 3 are viewed as supportive analyses while Analysis 1 is

		viewed as the primary analysis approach." • Section 2.1.7: Last sentence of second paragraph prior to Analysis 1 modified to: "These ANOVA tables are displayed in Appendix 2." • Section 2.1.8: Secondary: Under Partial AUC the following sentence was removed: " Differences will be absolute differences (pAUC(IUS)-pAUC(OA+IUS) with no normalization." • Section 2.1.8: Inter-reader variability: Analysis/Software: Inter-reader variability modified to read "Statistics will be calculated to assess inter-reader variabilities." • Section 2.1.9: First paragraph modified to "As for the PIONEER Pivotal Study, analyses will be based on independent readers for POM, BI-RADS and feature scores (IUS, Imagio (IUS+OA)) individually and overall. There is a formal hypothesis test for the primary endpoint and the three secondary endpoints. Hierarchical testing will be in place to control the Type I error. Truth (determination of malignant biopsy, benign biopsy, HR or TPB) will be taken from the PIONEER Pivotal Study. All CB masses (n=29) will be excluded from all analyses; these masses will be replaced in order to preserve final analysis." • Section 2.1.9: Language added on NLR, PLR and pAUC "NLR and PLR will be analyzed using the logarithmic transformation and the delta method. Calculations will be performed using MATLAB script: DLRATIOS.M. The changes in NLR and PLR correspond to the 10% specificity gain for the primary endpoint. Last, pAUC for fixed 95% to 100% sensitivity will also be analyzed using OR-BDM MRMC." • Section 2.1.9: Language removed on considering Chen R program • Section 2.1.9.3: Language to account for color balance added as well as new paragraph clarifying intent of interim analysis: "A blinded interim analysis is planned for this study after all readers complete the reads for the first three blocks (360 - 10 CB masses = 350 complete image sets). The CB masses (n=10) will be excluded from the interim analyses." The interim analysis will solely re-estimate sample size based on the primary endpoint; it will not be used to modify the alternative hypothesis. Neither futility nor efficacy will be assessed. The following recommendations will be issued to Seno by the unblinded statistician at the interim timepoint: 1. Sample Size Calculation: ___One additional block is required OR ___Two additional blocks are required OR ___Three additional blocks are required OR ___Four additional blocks are required OR

Version	Date	Description of Changes
		___Four additional blocks are not sufficient but reads should continue because there is a reasonable chance to reach statistical significance based on the conditional power assuming the alternative hypothesis holds. ___Four additional blocks are not sufficient, reads should not continue because there is not a reasonable chance to reach statistical significance based on the conditional power assuming the alternative hypothesis holds. • Section 2.1.9.4 was modified to: "To extend labeling claims beyond the primary endpoint, a hierarchical testing strategy will be used to control Type 1 error associated with testing the primary endpoint and three secondary endpoints in the following pre-defined order until statistical significance is no longer reached in support of extended labeling: • Increase in specificity for fixed 98% sensitivity. • NLR • PLR • pAUC over 95% to 100% sensitivity No other endpoints are to be included as labeling claims. • Section 3: The following sentence was removed: "If the Chen R program gives the same answers as the Hillis software, then we will consider using the software at the time of the interim analysis." • Section References: Chen reference number 1 was removed: "1. Chen W, Huang Z, Samuelson F, Tcheuko L (2019). Adaptive sample size reestimation in MRMC studies, SPIE 10952, Medical Imaging: Image Perception, Observer Performance, and Technology Assessment, 109520G" • Section References: Reference 10 was removed: 10. Statistical Assessment Methodology for Diagnostics and Biomarkers: https://www.fda.gov/medicaldevices/ cdrh-research-Programs/statistical-assessment-methodology-diagnostics-and-biomarkers https://github.com/DIDSR/iMRMC/releases • Various administrative changes END V4.0

V5	04Nov2019	START V5.0 • Section 2.1.1: Number 19 Modified to read: A hierarchical testing strategy will be implemented to extend labeling by testing effectiveness in the following order: gain in specificity for fixed 98% sensitivity, decrease in NLR, gain in PLR, and gain in pAUC over 95% to 100% sensitivity where gain is the difference between the OA/US and IUS values. • Section 2.1.4: Sample Size Calculations: Added the entire section including the results for the variance estimates • Section 2.1.10: Hypotheses to be Tested: Added entire section for clarification • Section 2.1.11: Data Analysis: Added language: "In clinical practice, the POM Scores (0-100) are standardized such that a BI-RADS 3 (probably benign) shows a likelihood of cancer to be > 0% up to 2% inclusive." • Section 2.1.11: Data Analysis: Added language: "Specificity and sensitivity are calculated using this fixed 2% cutoff. Please be assured that the OR-DBM MRMC 2.51 (Hillis) software does not pick or use a fixed cutoff. In our application, the Hillis software computes AUC, partial AUC, specificity gain at fixed sensitivity, and sensitivity gain at fixed specificity using a full factorial ANOVA incorporating terms for readers, masses, and modalities. The gain in specificity can be computed at a wide range of sensitivities. The 98% sensitivity corresponds to a 2% false negative (FN) rate, but this unrelated to the use of a fixed 2% POM cutoff for clinical decision making." • Section 2.1.11: Data Analysis: Added language: "The DLRATIOS routine computes the standard deviation for the ratios of two independent random variables corresponding to the sensitivity and specificity which are both used to compute NLR and DLR" • Section 2.1.11.3: Interim Analyses and Data Monitoring: Added the entire paragraph describing the Plan, Methodology, the sub title Timing and the paragraph on Execution. Also, under Sample Size Calculations: added number of masses to clarify blocks. • Section 2.1.11.3 Interim Analyses and Data Monitoring: Methodology: The following sentence was deleted "Thus, the alpha spend for conducting the planned interim analysis is small." And modified to: "They specifically state that "for conditional power within a range of 0.37 and 0.80, the sample size can be increased while leaving the critical test value unchanged"; this is the fundamental premise of the promising zone. Thus, the alpha spend for conducting the planned interim analysis is small." • References: Added the below references:

Version	Date	Description of Changes
		Huang Z, Samuelson F, Tcheuko T, Chen W. Adaptive Design in Multi-reader Multi-case Clinical Trials of Imaging Devices. Statistical Methods in Medical Research, 0(0) 1-20 Adaptive Designs for Clinical Trials. September 2018. Wittes J, Brittain E. The Role of Internal Pilot Studies in Increasing the Efficiency of Clinical Trials. Statistics in Medicine, 9, 65-72, 1990. Kieser, M., Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Statistics in Medicine 22:3571-3581. Friede T, Kieser M. Sample Size Recalculation in Internal Pilot Study Designs: A Review. Biometrical Journal. 48, 4, 537-555, 2006. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994 Dec; 50 (4): 1029-41. END V5.0