Statistical Analysis Plan

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

Philip Lavin PhD, FASA, FRAPS
Roger Aitchison

For

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

CONFIDENTIALITY STATEMENT

Statistical Analysis Plan

Signature Page

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

Reader Study-01

Written by:

Roger Aitchison
Roger Aitchison
Biostatistician
Boston Biostatistics Research Foundation, Inc.

09NOV2018
Date

Philip Lavin
Philip Lavin PhD, FASA, FRAPS
Chief Biostatistician
Boston Biostatistics Research Foundation Inc

9 Nov 2018
Date

Approved by:

A. Thomas Stavros, MD
Chief Medical Officer
Seno Medical Instruments, Inc.

9-Nov-2018
Date

Ann Waterhouse
Ann Waterhouse
Vice President QA and RA
Seno Medical Instruments, Inc.

12 Nov 2018
Date

Shaan Schaeffer
Shaan Schaeffer
VP Clinical Operations
Seno Medical Instruments, Inc.

9 Nov 2018
Date

Table of Contents

1. 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.

1.1. Statistical and Analytical Plans

1.1.1. Definitions and General Considerations for Data Analysis

1. All calculations will be performed in a GCP-compliant environment using SAS statistical software, version 9.2 or later, StatXact, version 10 or later or OR-DBM MRMC (2.5) or later.
2. This Statistical Analysis Plan (SAP) is based on Clinical Trial Protocol Imagio Feasibility MRMC Study of Optoacoustic Images vs. Imagio Ultrasound to Guide Decision to Biopsy, September 24November 1, 2018 – Version 3.0.
3. This study is a single arm, controlled, blinded, multi-reader, multi-case (MRMC) study using a sequential design.
4. This feasibility study will be registered at clinicaltrials.gov.
5. There will be up to 12 independent readers to evaluate images using standardized equipment at American College of Radiology (ACR).
6. All readers will read all images utilizing (IUS) and before (IUS+OA pre) and after the SenoGram (SG) prompt (IUS+OA post); prior to all image reads, all readers will have access to mammography images and a clinical history as collected in PIONEER.
7. A total of 155 complete read sets from the original PIONEER Intention-to Diagnose (ITD) population will be randomly selected from within the PIONEER Study. Only once mass is to be selected per PIONEER subject.
    - Cohort 1 will consist of 75 benign and 45 malignant masses classified by CDU as BI-RADS 3 to 5 selected at random in proportion to the original distribution of Bi-RADS classifications among subjects in the PIONEER Study. To facilitate the alignment of the PIONEER Pivotal Study data with the Feasibility Study data, the mass image set sampling plan will select a proportion with and without mammograms for benign masses. [Note: mammograms were present for nearly all malignant masses.]
    - Cohort 2 will consist of 30 malignant masses with at least one false negative read from the PIONEER study to assess if the excess FN rate observed with OA in the PIONEER study can be reversed. The same proportional sampling strategy with respect to CDU Bi-RADS will be used. Cohort 3 will consist of five DCIS, lymphoma, or phyllodes masses; this will assess if these atypical cases do not introduce FNs or FPs. [Note: Due to the small size, Cohort 3 results will be listed only and not be summarized.]
8. Cohort data will not be combined across cohorts.

9. To assess intermediate performance an interim analysis is planned after each reader has completed at least 72 Cohort 1 masses.
10. The primary endpoint is to evaluate the gain in specificity at fixed sensitivity for IUS+OA post vs IUS; IUS+OA pre-will be compared to IUS and IUS+OA post.
11. There are no formal hypothesis tests for this Feasibility Study; this study will be used to inform Seno on the design and sample size of a subsequent pivotal study.
12. Feasibility Study data will not be used for registration purposes.
13. Feasibility Study data will not be combined with the subsequent pivotal study but will be used to determine effect sizes to power the subsequent pivotal study.

1.1.2. Classification of Cancer Status

Feasibility Study masses will be classified as biopsy-confirmed cancer, biopsy-confirmed benign or Truth Panel benign from PIONEER.

1.1.3. Analysis Populations

Safety results for the safety population were reported for the entire sampling frame for this Feasibility Study in the PIONEER CSR. There will be no safety reporting for this study.

Masses will be selected according to the SENO Sampling Plan (11SEP2018).

There will be three cohorts as described above:

1. Cohort 1: PIONEER Feasibility Simple Random Sample
2. Cohort 2: PIONEER False Negatives (FN)
3. Cohort 3: PIONEER DCIS, lymphoma or phyllodes tumors

At study completion the number and type of protocol deviations (PD) will be reviewed to determine whether there are missing or incomplete reads that would impact analyses. Significant findings will report findings to the study sponsor and used to improve the pivotal study design.

1.1.4. Disposition of Masses

A total of 155 masses were selected from the PIONEER ITD population; 120 masses for Cohort 1, 30 for Cohort 2, and 5 for Cohort 3. A 10% over-sampling allowed for replacement of any mass that might prove unreadable for Feasibility. Replacement of any masses will be documented in the CSR.

1.1.5. Demographic and Other Baseline Characteristics

Clinical presentation and medical history (from the PIONEER database) of site-determined CDU Bi-RAD score, 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. Mass size and depth, as recorded by the feasibility readers, will be summarized using descriptive statistics.

Summary tables will display demographic and baseline characteristics by cohort and by diagnostic status (benign+TPB, cancer). Cohorts will not be combined for any displays or analyses.

1.1.6. Medical History and Concomitant Medications

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

1.1.7. Effectiveness Analysis

There are no formal tests of hypotheses for this study. P-values that are generated will be regarded as descriptive statistics. IUS+OA pre-is OA before the SenoGram; IUS+OA post is the reader input OA after viewing the SenoGram. The following will be used to assess feasibility of a pivotal study.

Table 1: Effectiveness Analysis for Study Endpoints

Endpoints by Cohort	Analysis/Software
Cohort 1: Primary	
Gain in specificity IUS+OA post vs. IUS at fixed sensitivity (95-99%) interpolated from the ROC AUC curves. This analysis will be repeated for IUS+OA pre vs. IUS and IUS+OA post vs. IUS+OA pre. A recommended pivotal study sample size will be determined using multiple sources, including OR-DBM MRMC sample size program. Review of entire feasibility study report will be used to recommend efficacy endpoints for the pivotal study.	OR-DBM MRMC Sample Size: OR-DBM-MRMC Power program, nQuery v7, SAS (PROC POWER), and/or PASS v15.05.
Cohort 1: Secondary	
Partial AUC: Difference (IUS+OA post vs. IUS) in partial ROC AUC curves for 95-99% sensitivity. This analysis will be repeated for IUS+OA pre vs. IUS and IUS+OA post vs. IUS+OA pre.	OR-DBM MRMC & methods as per McClish
Specificity and Sensitivity: Observed and model-adjusted specificity and sensitivity will be reported for IUS+OA post vs. IUS using a 2% POM cutoff. These analyses will be repeated for	SAS PROC GEE and MIXED treating readers as independent observations and readers as correlated and then readers correlated and masses as independent.

IUS+OA pre vs IUS and IUS+OA post vs IUS+OA pre, and for POM cutoffs of 1%, 3%, 4%, and 5%.	Two-sided 95% confidence intervals for sensitivity and specificity will be computed for IUS, IUS+OA pre, the IUS+OA post, and the 3 pairwise differences.
Negative Likelihood Ratios (NLR) for IUS and IUS+OA post, NLR defined as ((1-observed sensitivity)/observed specificity)	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.
BI-RADS (BR): Downgrade (IUS+OA post) defined as: • IUS in BR 4a and IUS+OA post in BR 2-3 or • IUS in BR 3 and IUS+OA post in BR 2 BI-RADS: Upgrade defined as: • IUS in BR 2 and IUS+OA post in BR 3-5 or • IUS in BR 3 and IUS+OA post in BR 4-5 Tables will be repeated for IUS x IUS+OA post, IUS vs IUS+OA pre and IUS+OA pre vs IUS+OA post.	SAS: BI-RADS Grade Changes (Grouped BI-RADS), simple counts with 95% CI. Benign+TPB: Reference PIONEER Table 14.2.13.4 [t_14_02_13_birch1_04] Cancers: Reference PIONEER Table 14.2.13.7 [t_14_02_13_birch2_07] Note: change from 99% CI to 95% CI
Net grade change – All masses – weight (W) that results in a significant OA+IUS vs IUS reflecting the FP advantage minus the weighted FN disadvantage.	SAS: PROC GENMOD and MIXED linear models will contrast difference in FN and FP to demonstrate clinical benefit. Tables to be repeated for IUS x IUS+OA post, IUS vs IUS+OA pre and IUS+OA pre vs IUS+OA post
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 separate analysis for each study cohort and with subgroup analysis for masses with and without mammogram BI-RADS data.	For each metric, a summary table for each reader (where appropriate) and for all readers to contain the following: • Measured value using Reader-01 feature data. • Estimated value from SenoGram cross-validation with PIONEER feature data. • Estimated value from SenoGram predictions for cohort masses with

b) Specificity at fixed sensitivity (98%) for SenoGram classification as in (a). c) Partial AUC (over sensitivity range 95% to 99%, inclusive, using the McClish normalization) for SenoGram classification as in (a). d) Partial AUC (over sensitivity range 95% to 99%, inclusive, using the McClish normalization) for SenoGram classification as in (a). e) Sensitivity and specificity for SenoGram classification based on prediction intervals and the SenoGram threshold. The prediction will be considered inconclusive if the SenoGram threshold lies within the prediction interval; otherwise the interval prediction is treated the same as a probability prediction. As in (a), each cohort will be analyzed separately, with subgroup analysis for masses with and without mammogram BI-RADS data. f) Specificity at fixed sensitivity (98%) for SenoGram classification using the prediction interval as in (d). Partial AUC (over sensitivity range 95% to 99%, inclusive, using the McClish normalization) for SenoGram classification using the prediction interval as in (d).	PIONEER data.
Cohort 1: Exploratory	
OA Feature Score Distributions: • Benign vs. malignant masses • Benign masses only – mammogram vs no mammogram POM Score Distributions:	Summary Statistics by reader. Average over readers and then summarize means for overall. Use Wilcoxon for test statistic. Reference PIONEER T14.2.10.1 [t_14_02_1x_feat-10-_01]

• Benign vs. malignant masses • Benign masses only – mammogram vs no mammogram	Reference PIONEER Table 14.2.9.1 [t_14_02_09_pom-01]
Cohort 2: Secondary	
Observed and model-adjusted sensitivity	SAS: PROC GEE and MIXED
BI-RADS: Up and downgrades (defined above) for cancers only	SAS: BI-RADS Grade Changes (Grouped BI-RADS), simple counts with two-sided 95% CI. Cancers: Reference PIONEER Table 14.2.13.7 [t_14_02_13_birch2_07] Note: change from 99% CI to 95% CI Tables to be repeated for IUS x IUS+OA post, IUS vs IUS+OA pre and IUS+OA pre vs IUS+OA post
Cohort 3: Listings Only	
All Cohorts (Combined) Summary statistics to describe the SenoGram usage [n(%)] by individual reader and overall. Summary statistics for the SenoGram effect on reader confidence.	SAS: Proc Freq & Proc univariate

1.1.7.1. Feasibility Analysis Methodology

Data Analyses

Hypotheses generating analyses will be based on independent readers for POM, BI-RADS and feature scores (IUS, IUS+OA pre, IUS+OA post) individually and overall. There are no formal hypothesis tests; p-values from all analyses are descriptive only. There are no adjustments for multiple testing. There will be no adjustment in p-values reported for the planned interim analysis after a minimum of 72 reads by each reader in Cohort 1. Truth (determination of malignant biopsy, benign biopsy, or TPB) will be taken from PIONEER.

The analyses below will be done for Cohort 1, repeated for malignant masses only for cohort 2. No summaries (listings only) will be done for the 5 masses in Cohort 3.

The specificity advantage of IUS+OA post vs IUS, IUS+OA pre vs IUS and IUS+OA post vs IUS+OA pre-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, IUS+OA pre IUS+OA post, and the 3 pairwise differences [PROC GENMOD, readers as independent,

repeated as masses as independent]. Specificity and sensitivity will be calculated (not model-adjusted) using the 2% POM cutoff for IUS+OA pre, IUS+OA post and IUS; and then for robustness analyses, for 1%, 3%, 4%, and 5% cutoffs.

Clinically relevant up and downgrades based on BI-RADs are defined as:

- Downgrades: IUS in BR 4a and IUS+OA post in BR 2-3 or IUS in BR 3 and IUS+OA post in BR 2
- Upgrades: IUS in BR 2 and IUS+OA post in BR 3-5 or IUS in BR 3 and IUS+OA post in BR 4-5.

These definitions likewise hold for differences for IUS vs IUS+OA pre and IUS+OA pre and IUS+OA post. These clinically relevant up and downgrades will be reported by reader and overall for benign and cancers separately. For benign masses, up and downgrades will be reported for masses with and without mammograms.

To assess the net effect of downgrades and upgrades, a net gain will be computed in SAS using contrast statements for the ANOVA to determine weight (W) of the malignant outcomes (TP and FN) to determine the value (A-B+W(C-D)) that achieves statistical significance, where A, B, C and D are defined as:

A. Benign Masses correctly downgraded using IUS+OA post
B. Benign Masses incorrectly upgraded using IUS+OA post
C. Malignant masses correctly upgraded using IUS+OA post
D. Malignant masses incorrectly downgraded using IUS+OA post

These analyses will be repeated for IUS vs IUS+OA pre, IUS vs IUS+OA post and IUS+OA pre vs IUS+OA post.

The sample size for a pivotal study will be calculated using method implemented in the MRMC Sample Size Program 1.0 for Diagnostic Studies, by Hillis, Obuchowski, and Birnbaum [7]. These sample size results may be confirmed by other packages such as nQuery, PASS, and SAS [PROC POWER].

POM summary statistics for IUS, IUS+OA pre and IUS+OA post, and the 3 pairwise differences will be displayed by diagnostic category and compared using a Wilcoxon statistic. 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.

1.1.7.1.1. Adjustments for Covariates

No adjustments for covariates are planned.

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

1.1.7.1.2. Handling of Dropouts or Missing Data

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

1.1.7.1.3. Interim Analyses and Data Monitoring

An interim analysis is planned for this study after a minimum of 72 reads by each reader to assess reader performance in Cohort 1.

1.1.7.1.4. Multiple Comparisons / Multiplicity

There is no adjustment for multiplicity for this study.

1.1.7.1.5. Examination of Subgroups

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

1.1.7.2. 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:

- IUS+OA post vs IUS+OA pre-as described in Table 1 - Cohort 1: Primary and Cohort 1: Secondary.
- N (%) masses with reader reporting use of the SenoGram.
- 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.

1.1.8. Safety Evaluation

1.1.8.1. Extent of Exposure

This is a reader study only. OA Exposure for these subjects was reported in PIONEER CSR. There is no additional subject exposure to report.

1.1.8.2. Adverse Events

This is a reader study. Adverse events were reported in PIONEER CSR.

There is no additional OA exposure.

1.1.8.3. 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.

1.2. Hypotheses to be Tested

No hypotheses are tested in this feasibility study.

1.3. Determination of Sample Size

This Feasibility Study was not powered to detect a specified improvement in OA performance. An objective of this study is to suggest sample size for a subsequent pivotal reader study.

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

The protocol states that the presence or absence of a mammogram will be incorporated into the sampling plan; presence or absence of a mammogram could be incorporated only for the benign+TPB masses, as there were too few (~3%) missing mammograms in the malignant masses.

REFERENCES:

1.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.

2. 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.

3. 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.

4. 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.

5. Hillis SL, Berbaum KS, and Metz CE. (2008). Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Academic Radiology, 15,

647-661. DOI:10.1016/j.acra.2007.12.015

6. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multi-reader ROC methods: An updated and unified approach. Academic Radiology 2011; 18:129-142. doi: DOI: 10.1016/j.acra.2010.09.007.

7. Multi-Reader Sample Size Program for Diagnostic Studies, available at http://perception.radiology.uiowa.edu/Software/PowerandSampleSizeEstimation/tabid/252/Default.aspx.