Project title:

# Evaluation of Computer-Aided Lung Nodule Detection Software in Thoracic CT for Riverain Technologies LLC

NCT number: NCT 02440139

Date: August 24th, 2017 (VT IRB # 15-345 Amendment Application Approval date)

## Subject:

Study Protocol and Statistical Analysis Plan - only required with results information for studies with a Primary Completion Date on or after January 18, 2017

## Introduction:

The document is created to fulfill the requirement indicated by ClinicalTrial.gov as an item of "Document Section" . We only have one version of Study Protocol and Statistical Analysis Plan as shown below if necessary.

## 1.0 STUDY OBJECTIVE

The primary objective of this clinical study is to prove that a user aided with ClearRead CT InSight is ***superior to the unaided reader for detecting actionable lung nodules.***

The secondary objective of this clinical study is to prove that the reader's reading time is not significantly increased when aided with ClearRead CT Insight.

## 2.0 STUDY ENDPOINTS

1. The scores, likelihood of cancer, along with clinical actions at specific cut-points given by the radiologists with and without ClearRead CT Insight and compared to the true status of the study-case. The frequency of the scores for each method (Unaided and aided with CLEARREAD CT INSIGHT) will be recorded. This information will be used to construct a localized receiver operating characteristic (LROC) curves. Cut points will be used to establish point based estimates, namely; Sensitivity, Specificity, PPV and NPV.

2. Reading times associated with unaided and aided reads will also be assessed.

3. Machine performance will be measured as free response operating curve (FROC) or nodule sensitivity versus patient false positive rate

## 3.0 STUDY DESIGN

### 3.1 OVERVIEW

A reader study with at least ten (10) participating radiologists will be conducted. A localized receiver operating characteristic curve will be used to evaluate radiologists' diagnostic performance (in terms of the trade-off between the sensitivity and specificity when the decision criteria changes) in the detection of lung nodules on lung CTs with and without the

usage of the *ClearRead CT Insight* system. The time needed for the review and interpretation of each case will also be recorded.

An initial (baseline) interpretation will be made by each of the radiologists based on the Lung CT in its original form. At a minimum of one month later, each radiologists will again interpret the same images viewing the pair of CRCTI CT series: Two sets of CT images (standard with CADe marks and processed with vessel suppression will be presented on either one large monitor or two adjacent monitors.

During the baseline reading the radiologist will mark the location of the actionable nodules and assign a score. The radiologist will also indicate the recommended method of follow-up (Contrast CT, PET-CT, CT Follow-up, Biopsy).

During the second reading session (concurrent read), the radiologist will be presented with a standard appearing CT with CADe marks placed (left image) and the vessel suppressed same slice with the vessel suppressed view (right image). The second image, vessel suppressed, will be grayed out until the radiologist move the mouse to the second panel. The radiologist will mark locations. These may or may not correspond to the locations of the CAD markers. As before, the radiologist will assign a level of suspicious to each mark and indicate the need, if any, of an additional diagnostic action (Contrast CT, PET-CT, CT Follow-up, or Biopsy).

Based on the levels of suspicion for each nodule and the associated likelihood ratings, LROC curves will be constructed for both the baseline and the concurrent reads and the significance of any difference will be calculated. The recommendations for further action (Contrast CT, PET-CT, CT Follow-up, or Biopsy) will be used to calculate sensitivity and specificity, PPV and NPV.

**Number and types of cases:**

Retrospective lung CT image series from approximately 300 patients will be included in the study. Approximately one hundred (100) of the patients will have pathology confirmed cancers and approximately two hundred (200) of the patients will be CTs associated with normal patients. Also included as nodule images are those where the actionable nodule was not acted on at that time, but was detected and acted on based on a subsequent CT. These are the prior images where the nodule can be identified and its location is the same as on the "current" image confirmed by the radiologist expert panel (using a majority of three as the decision criterion).

The selected sample, randomly selected from a larger pool of CT cases will be enriched in the following way:

1. Lung nodules (cancers in this study) will be tumor size T1a (20 mm or less). The proportion of nodules 20 mm or less may be increased since this is where we expect the major impact of this software to be.
2. Non-Solid (ground glass) nodules will be added to the sample (based on availability) to determine the performance of the system on non-solid nodules. For this group, to have sufficient cases, we may have to include benign (non-malignant) non-solid nodules.
3. In this project, VIRGINIA TECH will perform a Machine Test of the ClearRead-CT-Insight algorithm followed by a reader performance evaluation study. Riverain will provide a system configured with the operating point set to be used for the reader studies and a configuration for an "open" system to be used for machine testing and FROC generation. Figure 1 shows the reader performance study consists of two arms; **Arm 1:** a baseline read (no secondary content) and **Arm 2:** concurrent, CAD augmented read.

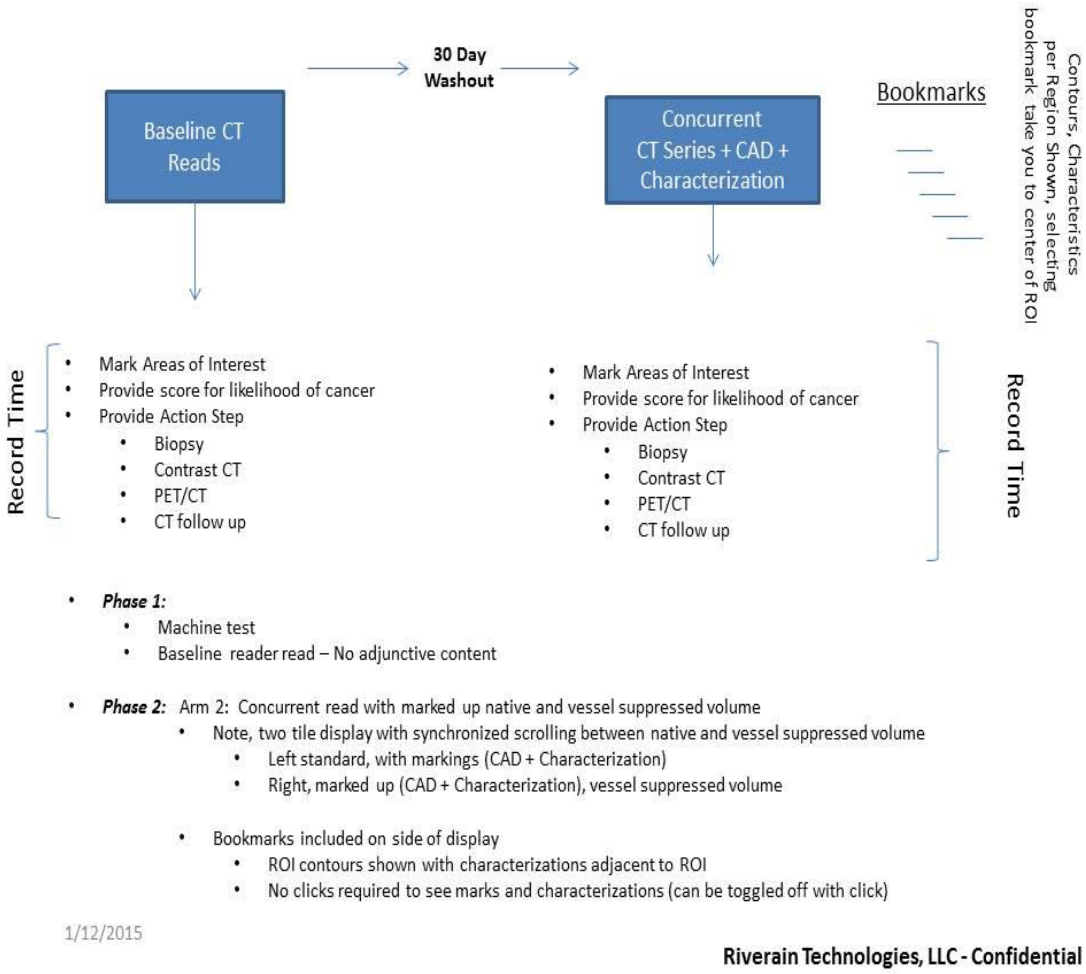


**Figure 1: ClearRead Insight Reader Study Design**

1st Arm: Do baseline (measure time, readers score regions according to action and suspiciousness) – mark all locations of concern

2nd Arm: Concurrent read (measure time, readers score regions according to action and suspiciousness) – mark all locations of concern

The primary study hypothesis is that the adjunctive use of ClearRead CT Insight is superior to use of standard lung CT images alone, as measured by the area under the LROC curve.

## 3.2 STATISTICAL ANALYSES

### *Accuracy*

To evaluate the hypothesis of superiority in terms of improvement in accuracy for ClearRead CT Insight vs. unaided, a mixed effects model (DBM) will be implemented (similar to the model outlined in Dorfman, Berbaum and Metz, 1997), where variance components will be included to account for reader, case, reader by case, reader by modality, case by modality and reader by case by modality. However, it is anticipated that the three-way interaction will be inestimable and will subsequently be dropped from the statistical model. Specifically, the hypothesis to test for superiority of ClearRead CT Insight vs. unaided is:

$$\mathbf{H_0: AUC_{unaided} - AUC_{ClearRead\ CT\ Insight} \geq 0.0\ vs.\ H_A: AUC_{unaided} - AUC_{ClearRead\ CT\ Insight} < 0}$$

The AUC of the LROC is the primary endpoint to evaluate accuracy and the test of interest will be a two-sided 95% confidence interval on the effect of modality (i.e. *ClearRead CT Insight minus unaided*). Significance will be concluded if the upper bound of the two-sided 95% confidence interval does not include zero. If the null hypothesis ($H_0$) is rejected, the alternative hypothesis ($H_A$) is accepted and the superiority of using the ClearRead CT Insight system will be established.

### *Time*

The second co-primary objective is to evaluate reduction in time spent per image for *ClearRead CT Insight vs. unaided*. Specifically, the hypothesis to test for superiority of *ClearRead CT Insight vs. unaided* is:

$$H_0 = T_{unaided} - T_{ClearRead\ CT\ Insight} \leq 0 \text{ vs. } H_A = T_{unaided} - T_{ClearRead\ CT\ Insight} > 0$$

To evaluate the hypothesis of non-inferiority in terms of improved read time for *ClearRead CT Insight* vs. *unaided,* a mixed effects model will be implemented (similar to the model outlined in Dorfman, Berbaum and Metz, 1992), where variance components will be included to account for reader, case, reader by case, reader by modality, case by modality and reader by case by modality. However, it is anticipated that the three-way interaction will be inestimable and will subsequently be dropped from the statistical model.

The read times will be tested using a two-sided 95% confidence interval on the effect of modality (i.e. *ClearRead minus unaided*). Significance will be concluded if the upper bound of the two-sided 95% confidence interval does not include zero.

With either analysis, the use of the mixed model could be modified to employ a bootstrap sampling approach if the model assumptions of the DBM method have been violated. The upper 95 % confidence limit for the difference in the areas under the curves would be calculated using 10,000 bootstraps in the MultiReader MultiCase (MRMC) ROC method.

## 3.3 POWER AND SAMPLE SIZE

The power to detect differences in the AUC of the LROC curve for the proposed statistical analysis using the current design baseline, i.e. 300 cases, each corresponding image read by 10 readers, was assessed through a simulation study. Specifically, the model outlined above was used to simulate 500 datasets across a range of effect sizes, where power was defined as the

proportion of datasets that yielded a significant p-value for testing the fixed effect of modality. The method of Dorfman, Berbaum and Metz, 1992 (DBM) was implemented utilizing statistical mixed model theory with jackknife estimates.

These simulations required assumptions regarding the magnitude of the variance components associated with the different random effects. As a pilot study to obtain estimates of variance components was not conducted, variance component estimates from Riverain study SoftView 510(k) (Record # BSSI-PR-09-00006) that uses a similar technology as the proposed ClearRead CT Insight device were used in all power simulations. Please note that the variance components listed in Table V were scaled by the total variance to represent the proportion of total variance explained by each component. All power estimates are dependent upon the appropriateness of the assumed variance components.

**Table V: Variance component estimates from previous Softview reader study**

| **Variance Component** | **Scaled Estimate** |
|---|---|
| Reader ($\sigma^2_{reader}$) | 0.0003 |
| Case ($\sigma^2_{case}$) | 0.1258 |
| Modality-by-reader ($\sigma^2_{mod*reader}$) | 3.8981E-05 |
| Modality-by-case ($\sigma^2_{mod*case}$) | 0.0257 |
| Reader-by-case ($\sigma^2_{reader*case}$) | 0.0127 |
| Residual ($\sigma^2_{res}$) | 0.8355 |

The mean difference $\Delta\mu = \mu_1 - \mu_2$ between modalities corresponding to a specified effect size $\gamma$ was then calculated as

$$\Delta\mu = \gamma * \sigma_{res}.$$

Effect sizes of 0, 0.2, 0.5, and 1 representing no effect, a small, a moderate, and a strong effect respectively, were evaluated in this simulation study with the following rationale:

a) An effect size of 0 was chosen to assess type I error in this study.
b) An effect size of 0.1 was chosen to correspond to a change in AUC of LROC of 0.084, which is 0.10 x $\sqrt{0.8355}$.
c) An effect size of 0.2 was chosen to simulate the expected small improvement in accuracy in time.
d) An effect size of 0.5 was chosen to simulate the expected moderate improvement in accuracy or time.

Table VI presents a summary of the power simulations along with additional scenarios with fewer cases with 8 or 10 readers. Superiority is assessed as the proportion of simulations in which the upper bound of the confidence interval on the AUC difference is less than zero. The other power assessment is on the statistical significance of the difference between modality.

**Table VI: Power estimates for superiority and statistically significant modality effect using previous Softview reader study to estimate variance components**

| | | | Power | |
|---|---|---|---|---|
| Number of Cases | Number of Readers | Effect Size | Superiority | Modality Effect |
| 300 | 10 | 0 | 0.086 | 0.082 |
| | | 0.05 | 0.606 | 0.490 |
| | | 0.1 | 0.978 | 0.974 |
| | | 0.2 | 1.0 | 1.0 |
| | | 0.5 | 1.0 | 1.0 |
| 270 | 10 | 0 | 0.060 | 0.060 |
| | | 0.2 | 1.0 | 1.0 |
| | | 0.5 | 1.0 | 1.0 |
| 270 | 8 | 0 | 0.058 | 0.080 |
| | | 0.1 | 0.932 | 0.878 |
| | | 0.2 | 1.0 | 1.0 |
| | | 0.5 | 1.0 | 1.0 |

The results show that the type I error is approximately controlled at the nominal level, i.e. 0.05, and therefore type I error is maintained at the desired level. Furthermore, the power to detect a change in AUC of at least 10% of the standard deviation is over 80% given the assumed magnitude of the variance components, even with only 270 cases and 8 readers.

Based on data from the literature referenced earlier in this report, potential variability and effect size for sensitivity can be assessed. Based on a mean difference in sensitivity of 0.0914 with overall standard deviation over all estimates of 0.1234 for an effect size of 0.0914/0.1234, or 0.741, the superiority of ClearRead CT Insight would be detected with adequate power.

There are several scenarios including reduced number of cases and a reduced number of readers. Even with 270 cases and only 8 readers, the power of the study should be sufficient to detect an effect size of 0.1.