Title: Quantitative Imaging Biomarker Alliance for Volumetric CT Image Analysis: Roadmap for a Staged Validation Plan

Long-Term Goals:

- To establish processes and profiles that will eventually lead to the acceptance of 3D volumetric CT by the field and regulatory agencies as (1) proof of biology, (2) proof of changes in pathophysiology, and (3) surrogate end-points for changes in the health status of patients.

Specific Aims:

- To standardize the quantification of volumes of anatomical structures, such as neoplastic masses, with x-ray computed tomography (CT).
- To identify and create coping strategies for all meaningful sources of variability in measurements of volume with CT, so that the "output" of each instrument precisely and accurately reflects the "input". See Appendix I, "QIBA 3D CT Matrix".

Context: Multi-stage, progressive, non-clinical and retrospective clinical trials with multiple milestones for making definitive go or no-go decisions. Work will be conducted under the aegis of the RSNA's (Radiological Society of North America) QIBA (Quantitative Imaging Biomarker Alliance), which is a consortium comprised of the FDA division of imaging research, the NCI, the NIST, the ACRIN (American College of Radiology Imaging Network), the major imaging equipment manufacturers (GE, Phillips, Siemens, etc.), the Extended PhRMA Imaging Group, and others.

Methods: Multiple stakeholders participating in accuracy and repeatability measurements by analyzing image sets and submitting the results for comparison by the QIBA according to pre-specified criteria.

Part I: Technical Characteristics. Quantification of test-retest intra-and inter-rater reliability.

A. Anthropomorphic Phantom: Images of complex shapes already acquired by the FDA/CDRH/OSEL.

Part I, Stage A1: Assessment of the image analysis technique. Intra-rater reliability, i.e., test-retest precision of measurement of a single image set by single image analysis operators (i.e., one image set, one image analysis software package, one image analyst per image analysis package). There will be no limit on the number of single image analysis techniques that may be field tested. Image analysis packages that meet quality criteria for precision by a single operator will progress to the next stage. See Appendix II, "Part I, Stages A1 and A2, Quality Criteria for Image Analysis Software for Phantom Data".

Part I, Stage A2: Further assessment of the image analysis technique. Inter-rater reliability of measurement of a single phantom image set by multiple image analysis operators using a single image analysis technique. Image analysis packages that meet quality criteria for inter-rater reliability will progress to the next stage. See Appendix II, "Part I, Stages A1 and A2, Quality Criteria for Image Analysis Software for Phantom Data".

Part I, Stage A3: Characterizing CT instrumental variability. Multiple image sets of the same phantoms re-scanned under "coffee break conditions". The initial data sets come from three scanners, and then expand to more instruments in an attempt to parse performance of the camera from performance of the image analysis methods. The goal is to ensure that the output for all cameras will be adequately precise and accurate when given the same input, i.e., when the same phantoms are scanned on different machines with comparable acquisition parametrs.

Note: "Ground truth" has already been established for each object in the phantom by physical measurements of its volume in "ex vivo".

Note: Some parallel processing is expected, particularly between Part I Stage A1 and A3.

B. Standard Clinical Data Set: A relatively small set of DICOM images of lung tumors already acquired by the NCI RIDER project and provided to the National Institute of Standards and Technology (NIST).

C. Standard Clinical Data Sets: Two somewhat larger sets of DICOM images of lung tumors already acquired by the NCI RIDER project selected as fit-for-first-purpose by the QIBA as follows:

Image Set 1: Small (1-to-5 mm): Smooth, well demarcated pulmonary nodules. These will be high resolution images (1-to-3 mm without gaps) of the type used to assess changes of small pulmonary nodules in diagnostic settings.

Image Set 2: Large (> 10 mm): For-registration, RECIST compatible, complex thoracic tumors which are both (1) isolated and (2) abutting normal structures or demonstrating other complex features (e.g., speculation). These will be "ordinary" images (5 mm without gaps) of the type most commonly encountered in global trials of investigational new drugs for patients with advanced disease.

Part I, Stage C1: Assessment of the image analysis technique. Intra-rater reliability, i.e., test-retest precision of measurement for pre-specified tumors (a.k.a. "marked up") by single image analysis operators (i.e., one image set, one image analysis software package, one image analyst per software package). There will be no limit on the number of single image analysis techniques that may be field tested, providing they meet design specifications for quality in Part I, Appendix I. Image analysis packages that meet quality criteria for precision by a single operator will progress to the next stage. See Appendix III, "Part I, Stages B & C, Parts I and II, Quality Criteria for Image Analysis Software of Clinical Data Sets".

Part I, Stage B2: Further assessment of the image analysis technique. Inter-rater reliability of measurement of a single phantom image set by multiple image analysis operators using a single image analysis technique. Image analysis packages that meet quality criteria for In Part I and Part IIa will progress to the next stage. See Appendix III, "Part I, Stages B & C, Parts I and II, Quality Criteria for Image Analysis Software of Clinical Data Sets".

Note: For each clinical data set, the results of all image analyses meeting quality criteria will be pooled to establish "ground truth" and the confidence intervals around the "true" volume of each pre-selected tumor in the RIDER data sets. Further qualification will be based on assessments of concordance between algorithmically derived changes in volume and manual assessments by experts in radiology who will use non-volumetric techniques.

Part II: Establish standards for using 3D volumetric imaging in a retrospective clinical trail.

A. Determine level of performance adequate for using 3D volumetric analysis in a clinical trial

Stage IIA1. The effect size required to classify a change in the volume of a small pulmonary nodule as malignant, i.e., the difference in volume between within subject measurements at Time 1 versus Time 2.

Stage IIA2: To quantify the effect size that is required to cross thresholds for a treatment-induced responses in categorical assessments, such as "Partial Response" and "Disease Progression".

B Determine appropriate imaging acquisition standards for use of 3D volumetric analysis

C Determine what type of evaluations are necessary to validate the use of 3D volumetric imaging

Part III: Diagnostic Accuracy. Begin with a single expert per software package who will work under ideal conditions with high resolution images. Use RIDER data sets to derive Kappa statistics, receiver-operator-characteristic (ROC) curves, likelihood ratios, etc.

A. Quantification of sensitivity and specificity in distinguishing categorical response variables, including Partial Response (PR), Stable Disease (SD), and Progressive Disease (PD).

<specific procedure, evaluation method…>

Data collection required

Markup requirements

Approach to using data

B. Correlation between 3D image analysis and "latent gold standard", i.e., RECIST

<specific procedure, evaluation method…>

Part IV: Progress to multiple image analysts.

<specific procedure, evaluation method…>

Part V: Progress to "real world" image resolution.

<specific procedure, evaluation method…>

Part VI: Efficacy & Effectiveness. Formal estimate of the value from 3D volumetric image analysis versus latent standard (RECIST) in terms of

A. Increased analytical power per subject,

B Length of time each subject needs to stay on trial, and

C Cycle time required to make critical GO or NO GO decisions about drugs.

The specific aim will be to compare time-dependent outcome measures based on RECIST to outcome measures based on volumetric analyses, such as time to response and progression free survival for (1) individual subjects, and (2) the sample as a whole before the trial concludes a drug is either effective or futile.