Volumetric CT in Lung Cancer:

An Example for the Qualification of Imaging as a Biomarker

Andrew J. Buckler, MS, P. David Mozley, MD, Lawrence Schwartz, MD, Nicholas Petrick, PhD, Michael McNitt-Gray, PhD, DABR, Charles Fenimore, PhD, Kevin O'Donnell, MASc, Wendy Hayes, Hyun J. Kim, PhD, Laurence Clarke, PhD, Daniel Sullivan, MD

Rationale and Objectives: New ways to understand biology as well as increasing interest in personalized treatments requires new capabilities for the assessment of therapy response. The lack of consensus methods and qualification evidence needed for large-scale multicenter trials, and in turn the standardization that allows them, are widely acknowledged to be the limiting factor in the deployment of qualified imaging biomarkers.

Materials and Methods: The Quantitative Imaging Biomarker Alliance is organized to establish a methodology whereby multiple stakeholders collaborate. It has charged the Volumetric Computed Tomography (CT) Technical Subcommittee with investigating the technical feasibility and clinical value of quantifying changes over time in either volume or other parameters as biomarkers. The group selected solid tumors of the chest in subjects with lung cancer as its first case in point. Success is defined as sufficiently rigorous improvements in CT-based outcome measures to allow individual patients in clinical settings to switch treatments sooner if they are no longer responding to their current regimens, and reduce the costs of evaluating investigational new drugs to treat lung cancer.

Results: The team has completed a systems engineering analysis, has begun a roadmap of experimental groundwork, documented profile claims and protocols, and documented a process for imaging biomarker qualification as a general paradigm for qualifying other imaging biomarkers as well.

Conclusion: This report addresses a procedural template for the qualification of quantitative imaging biomarkers. This mechanism is cost-effective for stakeholders while simultaneously advancing the public health by promoting the use of measures that prove effective.

Key Words: Quantitative imaging; therapy response; imaging biomarker; volumetric CT; regulatory pathway.

Efforts to develop public resources and open source tools for qualifying longitudinal volumetric computed tomography (CT) imaging as a biomarker were re-invigorated in 2005 by an informal alliance between the National Cancer Institute (NCI), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the National Institute of Standards and Technology (NIST) and the US Food and Drug Administration (FDA) (1–5). Preliminary work led to the organization of an inter-federal agency sponsored public workshop held at NIST headquarters in September 2006 (1). This workshop addressed the physical standards that would be required for qualifying medical imaging techniques as biomarkers. Stakeholders from academia, industry, regulatory agencies, patient advocacy groups, and scientific imaging societies participated. A model similar to the "Integrating the Healthcare Enterprise" (IHE) was endorsed as a means to organize and encourage collaboration among diverse stakeholders, given viable pathway for improving the success it has enjoyed. An alliance of this sort was thought to be necessary because the development of new or enhanced imaging technologies can be complex and expensive. Early phase justification of the costs, before commercial viability and medical value are established, can be difficult.

The Scientific Advisory Board of the Radiological Society of North America (RSNA) met in November 2006, and subsequently agreed to establish a "Quantitative Imaging Biomarker Alliance" (QIBA), modeled on the IHE process. One of the first three projects selected for piloting under the QIBA aegis was volumetric quantification at CT imaging.

Acad Radiol 2010; 17:107–115

From Buckler Biomedical LLC, Wenham, MA (A.J.B.); MRI and Computational Image Analysis Lab, Department of Radiology, Memorial Sloan-Kettering Cancer Center (L.S.); Division of Imaging and Applied Math Director, LAMIS Image Analysis Laboratory, FDA/CDRH/OSEL (N.P.); Radiological Sciences, David Geffen School of Medicine at UCLA (M.M.-G., H.J.K.); Information Technology Laboratory, National Institute of Standards and Technology (C.F.); R&D, System Solutions, Toshiba Medical Systems (K.O'D.); Imaging, Bristol-Meyers Squibb (W.H.); Imaging Technology Development Branch, Cancer Imaging Program (L.C.); Department of Radiology, Duke University Medical Center, and Science Advisor, Radiological Society of North America (D.S.); Volumetric CT Team, Quantitative Imaging Biomarker Alliance (Chair, A.J.B.; Extended PhRMA Imaging Group Chair and Co-Chair, P.D.M.; L.S., Co-Chair; N.P., M.M.-G., C.F., K.O'D., W.H., L.C., Members). Received November 28, 2008; accepted June 19, 2009. We would like to acknowledge the support of the NCI Cancer Imaging Program, its RIDER Project (http://grants.nih.gov/grants/guide/pa-files/PAR-08-225.html) and the CDRH/NIBIB Laboratory for the Assessment of Medical Imaging Systems (LAMIS) (http://imaging.cancer.gov/reportsandpublications/ReportsandPresentations/LungImaging/print). **Address correspondence to:** A.J.B. e-mail: andrew@bucklerbiomedical.com

doi:10.1016/j.acra.2009.06.019

This report outlines the initial emphasis of the Volumetric CT Technical Committee within QIBA (6). The potential benefits of the effort are described, the roles for each of the stakeholders is explained, and a staged approach for moving the process forward is explained.

OBJECTIVES

The goal of the QIBA Volumetric CT Technical Committee is to establish a methodology whereby multiple stakeholders collaborate to test hypotheses about the technical feasibility and the medical value of imaging biomarkers, specifically through the example of volumetric imaging. The working hypothesis is that volumetric imaging is an effective method for quantifying treatment-induced changes in tumor volume, and, ultimately, changes in the health status of patients with lung cancer. In this example as in others, the result would be an efficient means to collectively pursue qualification data accepted by regulatory bodies that can then be used by individual entities more cost effectively than if they had to pursue the qualification individually. After such data exist, they may be referenced without need for repetition in new drug applications by pharmaceutical companies and likewise in 510(k) registrations by imaging manufacturers. This mechanism is cost-effective for stakeholders while simultaneously advancing the public health by materially advancing the use of measures that prove effective.

Proceeding by way of example, the specific aims will include comparing time-dependent outcome measures based on clinical responses and unidimensional line lengths, such as those described by Response Evaluation Criteria in Solid Tumors (RECIST) (7), to analogous outcome measures based on volumetric analyses.

The expectation is that volumetric methods for assessing patients with cancer will be adopted if they are shown to be a better measure of clinical outcome and a more effective clinical trial tool. This "effectiveness" will be established by demonstrating that the added image analysis efforts are outweighed by the benefits of requiring fewer enrollees in clinical trials and shorter trial, shortening the duration of trials relative to the accepted alternatives, or provide better correlations with actual patient outcomes.

The initial emphasis will address the following objectives:

- Characterize the precision and accuracy of volumetric tumor measurements. This will be an essential prelude for understanding the threshold that will be needed to classify longitudinal changes in tumor volume as medically meaningful surrogates for changes in health status.
- Compare the sensitivity of volumetric measurements to RECIST/World Health Organization (conventional and modified) outcome measures. This will be necessary to determine if progressive disease can be detected significantly sooner with volumetric techniques than with uni- or bidimensional line-lengths placed on a single slice.

STEPS TAKEN TO ADVANCE THE FIELD

The Volumetric CT Technical Committee includes representatives from clinical practice, academia, professional imaging societies, regulatory agencies, government-sponsored scientific institutes, biopharmaceutical companies, image analysis software developers, and imaging device industry. Neither membership nor participation is restricted. All proceedings are open. The Committee will serve the public by acting as:

- Creators of performance profiles and roadmaps for qualifying volumetric measurements
- Stewards of objects and image sets for developing new image acquisition and analysis tools
- Archivists of performance metrics for existing tools

The Committee has performed a systems engineering analysis, and defined processes for identifying hypotheses to test. To test these hypotheses, several data sets have already been assembled. These data sets were contributed by the Reference Image Database to Evaluate Response (RIDER) (8,9) group from the NCI, the NIBIB, the FDA, and Memorial Sloan Kettering Cancer Center. These data sets include both CT patient examinations and CT scans of anthropomorphic phantoms. These data have been placed in a public web-accessible resource with plans to expand the data archiving effort under a new NCI Quantitative Imaging Network. The pharmaceutical companies and the imaging contract research organizations are assembling real clinical trial data in which patients have been followed until disease progression was documented. Each data set is described below in association with its corresponding stage of qualification.

THE IHE PROCESS

Sophisticated healthcare capabilities often depend on multiple systems from multiple vendors integrating properly. The IHE Initiative has, over the last 10 years, evolved a successful process for identifying necessary details of integration and facilitating their implementation and deployment on real products in real healthcare environments.

Although it oversimplifies the activities involved in establishing an effective, collaborative initiative with critical mass, the main steps in the process can be summarized as follows.

1. Clinical and technical experts identify and define a critical use case.
2. Technical experts create a detailed specification (called an IHE Profile) to address the use case, selecting and optimizing established standards.
3. Industry implements the specification in their systems.
4. IHE tests vendors' systems at carefully planned and supervised events called Connectathons.

An IHE Profile is a document that specifies claims and details. The claims tell a user what can be accomplished by following the Profile. The details tell a vendor what must be

implemented in their product before they can declare compliance with the Profile. The details may also define related user procedures necessary for the claims to be achieved.

QIBA intends to extend the IHE process. IHE focuses on engineering problems where the scientific problems have been solved thus avoiding areas with open scientific questions. QIBA, on the other hand, strives to identify and refine key scientific precursor questions and concepts needing validation. QIBA then coordinates the research and other groundwork to achieve resolution of the scientific questions so the Profile definition can proceed.

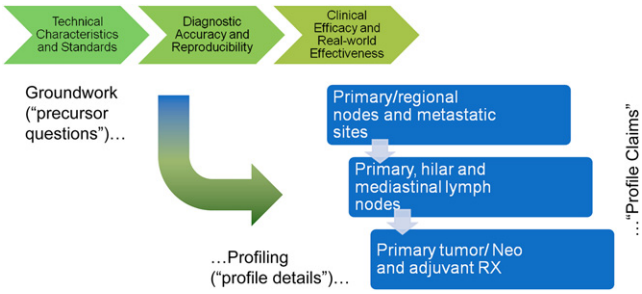

Figure 1. The pattern for one biomarker (lung nodules).

THE QIBA PROFILES FOR VOLUMETRIC CT

The QIBA CT Profiles are implementation specifications that detail the features, procedures, and interoperable data exchange required to achieve accurate, reproducible, clinically effective quantification. Research or clinical sites that follow the Profile procedures and use equipment that complies with the Profile can expect to achieve the capabilities the Profile claims.

By introducing Profiles in a step-by-step approach, incremental effort can provide incremental benefits. The first profile (applicable to advanced disease) has relatively few unknowns, is useful on its own, and, importantly, serves as a foundation which will accelerate the work to validate the later profiles (8).

Writing Profile documents requires first answering a number of "precursor questions." The first profile has relatively few, but subsequent profiles require answering more unknown/precursor questions. Fortunately, the groundwork research to answer many of these questions is already underway. The linkage of the precursor questions helps identify which of the ongoing groundwork projects to engage, support, or accelerate in order to specify/implement/validate the next profile. This chain establishes the common interest between researchers, imaging vendors, pharma, and physicians including radiologists and oncologists.

A set of three initial Profiles with progressively more sophisticated goals is illustrated in Figure 1.

Further detail on the approach, specific Profile claims, and other working material of the team is maintained on a Wiki page that enables the group activity (10).

After the first profiles are off the ground, we can shift focus to next target (eg, non-volumetric measures such as density or cross-modality mechanistic measurements) and work on the next set of profiles. The pattern could be repeated and tailored for other biomarkers (likely skipping through many of the steps much more quickly because many answers found for the lung cancer work will be the same or analogous).

THE QIBA VOLUMETRIC CT VALIDATION ROADMAP

The alliance has now built a staged analysis plan for validating volumetric CT as a biomarker. Briefly, the roadmap begins by characterizing the precision, accuracy, and stability of measurement in simple objects of known volume (11). A "go" decision to proceed to the next step is contingent on the acquisition and analysis meeting threshold requirements for quality. The successful analysis of simple geometric shapes leads to the analysis of complex shapes in anthropomorphic phantoms. The successful acquisition and analysis of complex phantom data leads to the analysis of small clinical data sets in which real lung tumors have relatively simple shapes and sharp contrast with surrounding lung tissue. Success leads to progressively more and more complex clinical data sets until the process is ultimately qualified for deployment in clinical practice and in multicenter trials. Conceptually, the roadmap is divided into discrete stages, which are described in the following section.

Part 1: Static Image Sets

Part 1A is under way and is analyzing bias and variance of various size measurement techniques where the nodule reference standard (ie, "ground truth") size is known deterministically. This stage addresses the question of what is the measurement variation under a limited set of controlled conditions for a reference set of phantom image data. QIBA is in the process of evaluating inter-/intra-reader reliability for the evaluation of anthropomorphic phantom data using a single commercial review and volumetric software package. The study protocol and reading procedure have been completed and a pilot study is being conducted to identify remaining issues that should be addressed before starting the pivotal study. The pivotal study has six radiologists using a commercial clinical review workstation to estimate the size of 40 phantom nodules. The readers estimate lesion size using volumetric, unidimensional, and bidimensional size estimation techniques. Each nodule is evaluated with all three techniques in each of two reading sessions, separated by a 3-week interval to allow for intra- and inter-reader variation estimates. The image data is from CT scans of the anthropomorphic phantom acquired during a separate FDA/NIBIB research project. The phantom nodule and CT acquisition characteristics for the study data include the following.

- Nodules characteristics:
- 10-mm and 20-mm solid spherical lesions

- 10-mm solid lobulated and spiculated lesions
- 20-mm solid ovoid lesions
- +100 HU and -10 HU densities for each nodule
- CT acquisition characteristics include
- 100 mAs exposures
- 0.75- and 5.0-mm slices
- 1 recon kernel

The nodule reference standard sizes have been estimated using "ex vivo" measurements based on micro-CT imaging and precise weight and materials density measurements for the individual nodules (Fig 2 and 3).

Part 1A is being conducted by a sub-team under the leadership of Nicholas Petrick and colleagues at the FDA/NIBIB as part of their intramural research program. Completion of the full 1A reader study is expected in late 2009.

Part 1B is using clinically acquired patient datasets to address issues relating to volume analysis of pulmonary nodules in diagnostic settings. It is investigating several issues regarding the question of what is the measurement variation for different software/user methods for a reference set of clinical image data. The first question to be investigated is to determine the level of accuracy and precision that can be achieved in measuring tumor volumes in patient datasets with a "known" volume. This investigation will utilize cases from the Lung Image Database Consortium data, which have nodules that have been contoured by as many as four readers and that will constitute the reference standard for the investigation. The second question will be to determine the minimum detectable level of change that can be achieved when measuring tumors in patient datasets under a "no change" condition (as illustrated in Figure 4). This "no change" condition is achieved by performing two CT exams on the same patient over the space of a few minutes and measuring tumor volumes from both exams. This investigation will be using a unique dataset provided to the RIDER database by investigators at Memorial Sloan Kettering Cancer Center. These experiments are being designed and data collection will begin in late 2009.

Further investigations will extend into patient datasets under conditions where the amount of change is not known. These investigations will utilize existing cases of patient change data from the RIDER study. The available RIDER CT data set we will be using consists of:

- 300 cases as of February 2009
- Most with multiple time points that are weeks to months apart (Fig 5)
- Few with thin slices (reconstruction intervals of <1.5 mm without gaps)
- No contours provided for the most part (there is a limited set of cases that have been annotated by two readers with RECIST markings; all on datasets with 5 mm reconstructed slice thickness)
- No outcome data
- Publicly available on NCIA (12)

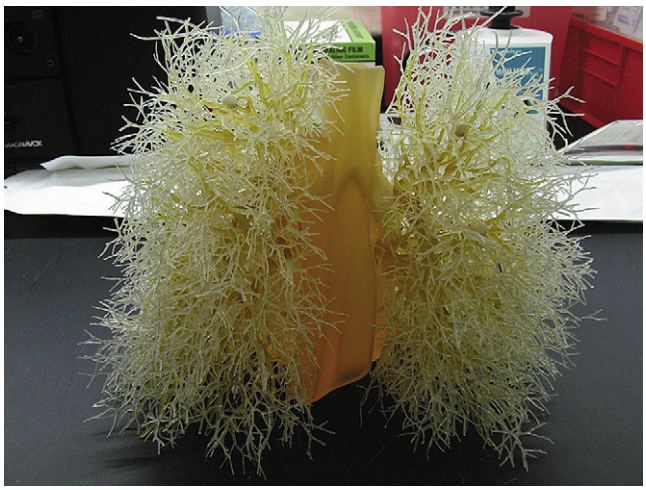

Figure 2. Photograph of the lung phantom being scanned by the Food and Drug Administration/National Institute of Biomedical Imaging and Bioengineering research team.

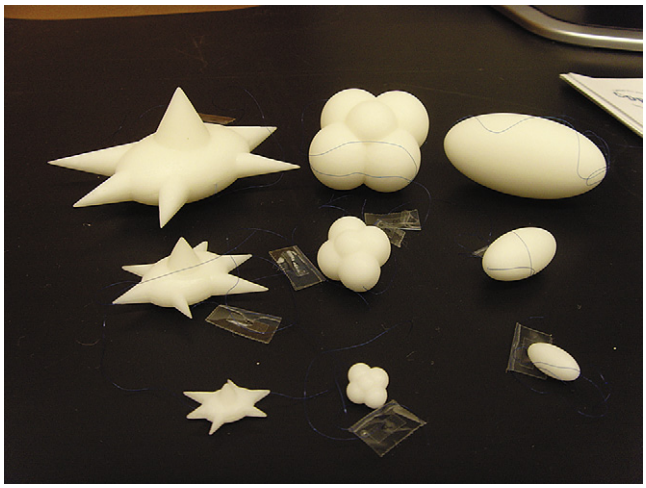

Figure 3. Examples of nodules that can be embedded in the phantom. The nodules are spiculated, lobulated, and ovoid, left to right, respectively. The phantom nodules shown have a volume equivalent to a sphere with a diameter of 40, 20, and 10 mm, top to bottom, respectively.

Data from RIDER will allow us to evaluate inter-reader variability in the measurement of change, for prespecified tumors (a.k.a. "marked up") by different image analysis operators (readers, algorithms, reader/algorithm combinations). In these cases, the reference standard volume change is unknown, so an analysis of reader variance (not accuracy) will be performed (Fig 4 and 5). Note: The RIDER project addresses both access to longitudinal and repeat images of phantom and patients (CT, positron emission tomography-CT, dynamic contrast enhanced-magnetic resonance imaging (DCE-MRI), diffusion-weighted imaging–magnetic resonance imaging). This work includes consensus discussions on how to perform analysis of repeat and longitudinal measurements and includes results of different measurements

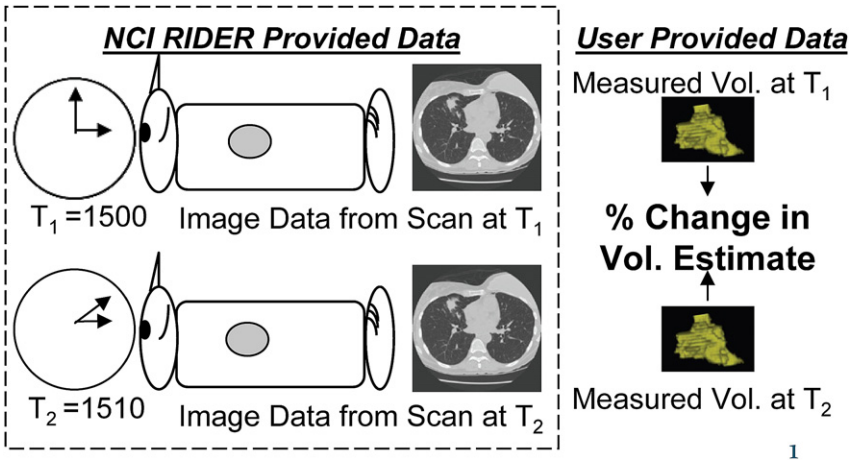

Figure 4. "Coffee break" conditions to evaluate measurement error when there is no biological change.

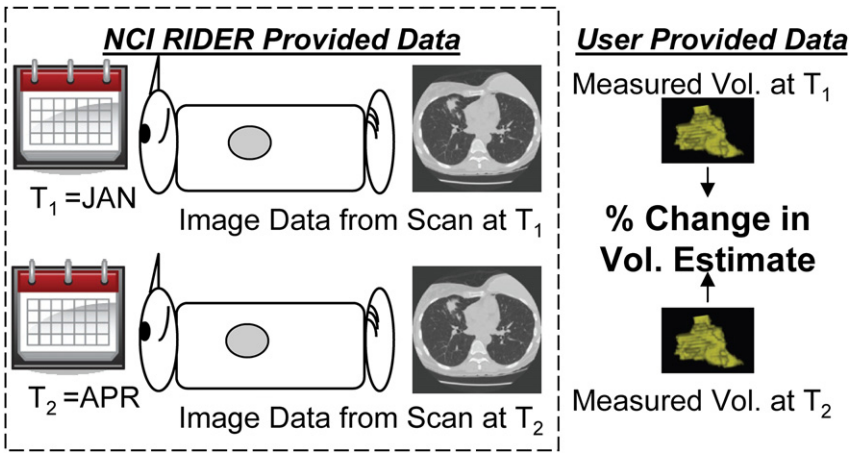

Figure 5. Change analysis when it is expected that biology may have changed.

is now available on NCIA as a public resource. A full report is to be published on this work by late 2009.

Part 1B is being conducted by a sub-team under Michael McNitt-Gray's leadership. Dr. McNitt-Gray has provided QIBA information on the consensus developed for CT methods of analysis from Lung Image Database Consortium and RIDER data over the last 2 years.

Part 1C is planning a multicenter, multiscanner phantom study of measured volume to model the sources of variability attributable to the scanner type and the center. The design parameters are selected to be relevant to one or more clinical scenarios. The study calls for collections of CT data using the lung phantom with multiple nodules described in Part 1A. A commercially available software tool will be used by each of several radiologists to analyze one-, two-, and three-dimensional (3D) (RECIST, WHO, and volume) features of the nodules. As noted in Part 1A, the size and shape of the phantom nodules have been estimated independently of the scan process, For this reason, the study can support the direct analysis of measurement *error* dependence on the platform, the center and the reader. The design is similar to that of inter-laboratory comparisons that are used to determine the reproducibility of complex measurement or manufacturing processes (13,14).

Multiple image sets of the same phantoms will be rescanned within and between centers to isolate contributors to variability. The goal is to determine necessary control conditions to be documented in QIBA Profiles. This will ensure that the output for imaging when performed under these conditions will meet defined precision and accuracy levels when scanned on profile-compliant equipment (Fig 6).

The primary basis for the activity is drawn from a systems analysis of the sources of variability in volumetric CT represented in a working matrix. Based on this analysis, sources of variability that are relevant to each considered profile constitute "factors," only some of which are can be varied. We have identified a number of factors important to our design including modality physics (particularly cone vs. linear beam), scanner design (including manufacturer, number of detector rows, and vendor specific system gain settings), acquisition protocol (field of view and settings of mAs and kVp), reconstruction (such as variation in kernel and algorithm). Most of these factors are fixed in our design to correspond to one of the clinical scenarios. The main factors to be varied systematically are the selection of the imaging center, the manufacturer, and the scanner type.

Part 1C is being conducted by a subteam led by Charles Fenimore. Dr. Fenimore has led the "Biochange 2008

Figure 6. The same phantom used in 1A is used for the study of system-dependent measurement errors (part 1C). This inter-center comparison study aims to characterize sources of variability complementary to those of 1A, including intra-machine effects (such as reconstruction filter) and inter-machine variability (such as manufacturer, number of detector rows, and imaging center).

TABLE 1. Sample Size for 3 Paired Methods at α = 0.05 and 80%+ Power as a Function of the Minimum Detectable Difference and Standard Deviation (SD) within the Group

	Minimum Detectable Difference (%)			
SD within group	1	5	10	20
1	51	4	3	2
5	335	51	15	6
10	335	188	51	15

TABLE 2. Sample Size for 5 Readers at α = 0.05 and 80%+ Power as a Function of the Minimum Detectable Difference and Standard Deviation (SD) within the Group

	Minimum Detectable Difference (%)			
SD within group	1	5	10	20
1	66	5	3	2
5	202	66	18	6
15	202	202	66	18

Benchmarking Pilot" (15), a study of lung CT change measurement algorithms and CAD tools conducted through the Information Technology Laboratory programs of the NIST.

Statistical Analyses

The previous studies all seek to estimate some measure of bias or variance. Some of these studies will be performed with an independently determined "truth" standard and compared to measurements derived from CT images; others will not have this independent truth standard and so only the variability between observers (readers, algorithms, or reader/algorithm combinations) will be assessed. For each study, sample size calculations will be performed so that the experiments will be properly sized to determine an effect at a specified level of power.

In general, two kinds of sample size calculations will be performed. When the focus is on comparison between either measurement methods or readers, then a multiple comparison test approach will be used. For example for the Part 1A experiment to evaluate inter- and intra-reader reliability with a single agreed upon software package, with analysis on anthropomorphic phantom data, the sample size can be calculated using a single factor analysis of variance study of multiple comparison test. Specifically, a Tukey-Kramer (pairwise) in the measurement of difference between readers (or method) at α = 0.05 and 80%+ power can be performed (16). Some example sample size calculations are shown in Tables 1 (for three methods) and 2 (for five readers).

When the focus is on evaluating the magnitude of the difference between methods or readers (and not just whether there is a statistically significant difference or not), then a binomial-based approach will be used. For these questions, measurements can be dichotomized using different thresholds, such as using different levels of agreement between image-based and independent measurement (say 1%, 5%, and 20% threshold for agreement with independent measurement). When this approach is used, then differences between volume measurement methods and volume from the independent measurement can be compared and sample size can be obtained from two-sided binomial test of differences in proportion for α = 0.05 and 80%+ power (17). These analyses will allow us to properly size the studies being carried out.

Part 2: Setting Standards for Using Volumetric Imaging by Retrospectively Reanalyzing Results from Clinical Trials

Part 1 of our staged approach to qualification is based on "static" images of phantoms or patients with lung cancer. Part 2 extends the process of testing hypotheses about longitudinal changes within subjects. Goals include the following.

A. Determine level of performance with respect to statistical power adequate for using 3D volumetric analysis in clinical trials
B. Determine appropriate imaging acquisition standards for use of 3D volumetric analysis
C. Determine what types of evaluations are necessary to validate the use of 3D volumetric imaging

Stated another way, the intent of Part 2 is to derive similar performance indicators under the "volumetric CT as a biomarker" as currently exists for RECIST. Figure 7 demonstrates the current reference standard under RECIST, and as such, what would be necessary to be comparable to RECIST.

Even if it is assumed that the example shown reflects the true pattern of change, the RECIST categorical response variables are still sometimes slow to make assessments of response, particularly when working with some targeted therapies. In the case shown, the tumor started growing several months before a diagnosis of progressive disease was triggered. The question is whether volumetric analysis can benefit individual patients and clinical trials by assessing progressive disease sooner than changes in unidimensional line-lengths (Fig 7).

Table 3 shows the relationships between line-lengths and perfect cubes and balls that uniformly contract or expand, demonstrating the promise of added sensitivity in a volume measurement even for the simplest shapes, where RECIST should be ideal.

The volume changes more than the line length, even in the simplest shape. Figure 8 emphasizes the point by illustrating the complex shape of malignant mesothelioma, where unidimensional line-lengths fail most conspicuously and a volume measurement would seem to be superior.

In this case, the line lengths have no apparent correlation to the size of the lesion, whereas volume would be expected to be more indicative of the biological burden.

DIAGNOSTIC ACCURACY AND REPRODUCIBILITY: ADDRESSED USING NEW STANDARDS

The plan progresses to applying the standard on select relevant patient CT data sets with high-quality images meeting standards established by preliminary phantom work and proven clinical outcomes, including overall patient survival.

III. Begin with a single expert per software package working under ideal conditions with high resolution images.
   A. Quantification of sensitivity and specificity for individual expert readers using prediction of survival at relevant established time-point (eg, 6-month survival for advanced lung cancer). This allows performing receiver-operating characteristic analysis of the sensitivity vs. specificity for different scoring approaches of cancer change in response to treatment. Cox hazard proportion measurements and Kaplan-Meier survival analysis can also be used to measure outcome prediction of different imaging methods.
   B. Correlation between 3D image analysis and latent standard, RECIST
IV. Progress to multiple image analysts

Metrics with high intra- and inter-observer agreement (eg, kappa statistics, other measures) are selected that generalize well across sites. The statistical assessment takes into account methods such as fitting Cox proportional hazards model for each variable separately and can then be applied to the time to progression for each group. Generalized R-square calculations are used to measure the predictive power for each of the measures.

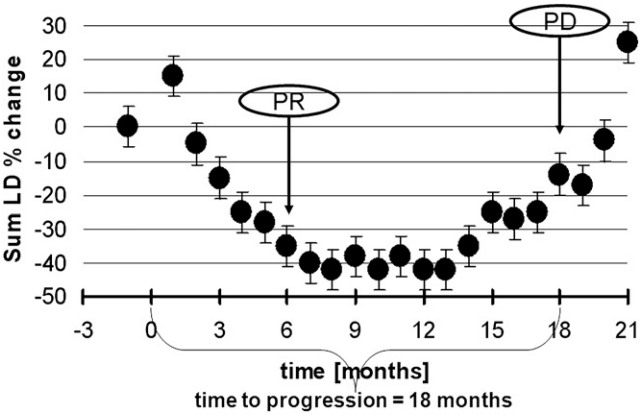

Figure 7. Plot showing time course with Response Evaluation Criteria in Solid Tumors definitions (11). In this example, a reduction in the sum of the longest diameters (SLD) of >30% is achieved at the 6-month mark and confirmed at the next time-point which is more than 4 weeks later, so the best overall response is partial response. An increase in the SLD of >20% greater than the nadir is passed at the 18-month mark, even though the absolute SLD is less than the baseline value, yielding a progression-free survival of 18 months.

END OF THE QUALIFICATION ROADMAP: ASSESSMENT OF CLINICAL EFFICACY AND VALUE

V. Progress previous analysis to "real-world" image resolution.
VI. 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 or a treatment regimen in purely clinical settings, and
   C. Cycle time required to make critical GO or NO GO decisions based on group differences between treatment arms in clinical trials.

Stages V and VI will explore the potential value of volumetric imaging in the general practice of medicine and in multicenter trials of investigational new drugs. The data sets will be contributed by pharmaceutical companies and their allied imaging contract research organizations. The CT scans will have been accrued during the conduct of multicenter trials, where the differences between treatment arms were statistically significant in terms of progression-free survival based on classical RECIST line-lengths.

The first data set will include 100 patients in each arm that were followed longitudinally until a diagnosis of progressive disease was made. The original image mark ups will be provided and used to select target lesions for volumetric analysis. This strategy of constraining the selection of target lesions should provide an "apples-to-apples" comparison of changes in volumes to changes in line-lengths.

Estimations will include how long it would have taken each individual subject to have been diagnosed with progressive disease if changes in tumor volume had been used as the basis for assessment. Descriptive statistics will be compiled for all the subjects in each arm. The length of time that would

TABLE 3. Relationships between Categorical Responses Based on Line-lengths and their Corresponding Changes in Volume

	Radius of Ball	Longest Line-Length of Cube	Δ Longest Line-Length Relative to Baseline	Δ Volume Relative to Baseline
Baseline	0.50	1.0		
Progressive disease	0.60	1.2	20%	72.8%
Partial response	0.35	0.7	-30%	-65.7%

Typical relationships between line-lengths and volume measures for uniformly contracting or expanding balls (solid spheres) and cubes with longest diameters of 1 unit in length. For a ball, the radius is one-half the diameter. For a perfect cube, the longest line is the hypotenuse of a right triangle on its surface.

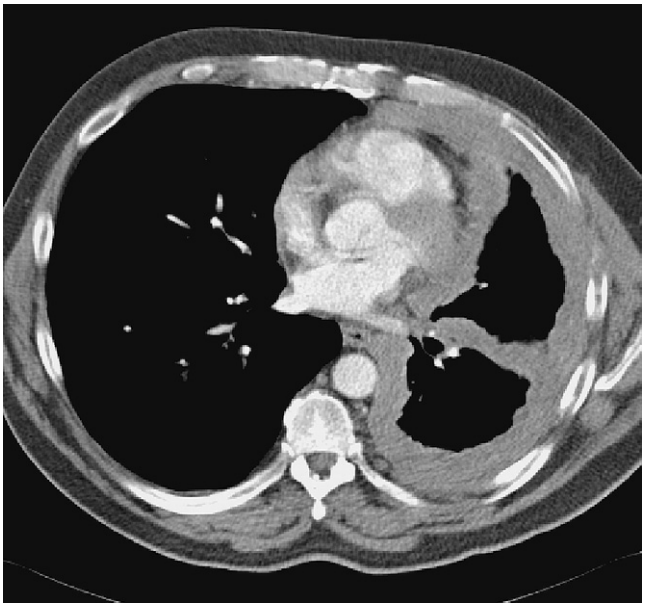

Figure 8. Some tumor shapes are not well modeled by line lengths.

have been required to see a difference between the two arms will be compared to the time that was actually required based on changes in line-lengths and the emergence of new tumor masses.

The long-term plan includes developing analogous data sets with a variety of characteristics, such as small nodules imaged with higher resolution parameters in neoadjuvant studies of patients with early-stage disease.

NEXT STEPS…

Part 1 is well under way, we have formed a sub-group for Part 2, and preliminary work on scoping the rest of the effort is being considered. On this basis, we intend to:

- Involve the broad clinical oncologic community (eg, American Society of Clinical Oncology (ASCO), European Organization for Research and Treatment of Cancer (EORTC), in specifying how a new imaging biomarker could be validated for measuring tumor response or progression in lung cancer as well as for other cancer sites.
- Set expectations regarding effectiveness and timing for principal objectives, supporting business models desired by pharmaceutical industry and imaging industry.
- Identify funding mechanisms, project plans, and reporting means for the identified activities.

The new NCI initiative (PAR U01) may provide funding and a research framework for some of this planned work. The planned cooperative group resulting from this new NCI initiative will have a steering committee together with an external advisory committee, and RSNA QIBA representatives will likely be members of those committees.

ACKNOWLEDGMENTS

The members of our Technical Committee who made substantial contributions to the work described in this report include: Denise Aberle of UCLA, Alaaddin Akkaya of Perceptive Informatics, Ricardo Avila of Kitware, Martin Barth of Definiens, John Boone of UCDavis and the American Association of Physicists in Medicine (AAPM); Andrew Buckler of Buckler Biomedical chairs the group and is the primary author of the matrix that attempts to describe all the sources of variability in measurements of volumes with CT, Charles Clark of NIST, Laurence Clarke of the NCI, which funded the creation of the RIDER database; David A. Clunie of RadPharm, Ekta Dharaiya of Philips Healthcare, Lori Dodd of NIH, Richard Eaton of MITA (NEMA), Charles Fenimore of NIST leads the Stage 1C subgroup, Robert Ford of RadPharm, Ronald Gottlieb of Roswell Park Cancer Center, David Gustafson of Intio, Wendy Hayes of Bristol-Meyers Squibb, Bruce J. Hillman of University of Virginia Health System, C. Carl Jaffe of Boston University, Philip F. Judy of NLST, Frank Klein of Definiens, Gerhard Kohl of Siemens Healthcare, Libero Marzella of FDA; Michael McNitt-Gray of UCLA leads the Stage 1B subgroup, P. David Mozley of Merck served as a cochair and is the lead author of the validation roadmap on our Wiki, James Mulshine of Rush is the primary driver of our "group 2" effort that focused on the medical utility of this work, Daniel Nicolson of Definiens, Morgan Nields of Intio, Kevin O'Donnell of Toshiba introduced the concepts of profiles and is the primary author of the texts describing

them on our Wiki, Scott Peairs of Intio, Nicholas Petrick of the FDA leads the Stage 1A subgroup, Sam Richard of Duke University Medical Center, Ehsan Samei of Duke University Medical Center, Larry Schwartz of Memorial Sloan Kettering Cancer Center served as a cochair, Elliot Siegel of U of Maryland, Sandra Scheib of Roswell Park Cancer Institute, Paul Licata of GE Healthcare, Daniel C. Sullivan of Duke University founded the group and serves as the scientific advisor to RSNA, Matthias Thorn of Siemens Healthcare, Binsheng Zhao of Memorial Sloan Kettering Cancer Center is the primary investigator who contributed the "coffee break" data set in patients with lung cancer, Linda Bresolin, Joseph Koudelik, and Fiona Miller of the RSNA organized each meeting, kept our records, and built the contents of our Wiki. There would have been nothing without them.

The lung phantom data being used in this QIBA effort were collected and provided as a public resource by the Office of Science and Engineering Laboratories within the US FDA (DIAM/OSEL/CDRH/FDA). We would like to acknowledge the FDA lab's efforts and particularly the efforts of Lisa M. Kinnard and Marios A. Gavrielides in the collection and distribution of this phantom data.

REFERENCES

1. Clarke L, Schilling LB, Sriram RD. Imaging as a biomarker: standards for change measurements in therapy workshop summary. NIST Interagency Rep 2008. NISTIR7434:1.
2. McLennan G, Clarke LP, Hohl R. Imaging as a biomarker for therapy response: cancer as a prototype for the creation of research resources. Clin Pharmacol Ther 2008; 84:433–436.
3. Armato S 3rd, Meyer C, McNitt-Gray M, et al. The Reference Image Database to Evaluate Response to Therapy in Lung Cancer (RIDER) project: a resource for the development of change-analysis software. Clin Pharmacol Ther 2008; 84:448–456.
4. Petrick N, Brown DG, Suleiman O, et al. Imaging as a tumor biomarker in oncology drug trials for lung cancer: the FDA perspective. Clin Pharmacol Ther 2008; 4:523–525.
5. Gavrielides MA, Kinnard LM, Myers KJ, et al. Noncalcified lung nodules: volumetric assessment with thoracic CT. Radiology 2009; 251:26–37.
6. Buckler A. Volumetric CT: a potential biomarker of response. Drug information agency (DIA) meeting held on 3 October 2008 "Medical imaging continuum: path forward for advancing the uses of medical imaging in the development of new biopharmaceutical products."
7. Eisenhauer E, Verweij J, Therasse P, eds. Response assessment in solid tumours (RECIST): version 1.1 and supporting papers. Eur J Cancer 2009; 45:(2) 225–310.
8. National Cancer Institute. Reference Image Database to Evaluate Response (RIDER). Available online at: http://gforge.nci.nih.gov/projects/rider/. Accessed October 24, 2009.
9. Qualitative Imaging Biomarkers Allicance. Volumetric CT. Available online at: http://qibawiki.rsna.org/index.php?title=Volumetric_CT. Accessed October 24, 2009.
10. Prionas ND, Ray S, Boone JM, et al. Volume assessment accuracy in computed tomography: a phantom study. Journal of Investigative Medicine 2009; 57:50–96.
11. National Cancer Institute. National Biomedical Imaging Archive. Available online at: https://imaging.nci.nih.gov/ncia/. Accessed October 24, 2009.
12. National Institute of Standards and Technology. Weights and measures inter-laboratory comparisons. Available online at: http://ts.nist.gov/WeightsAndMeasures/Metrology/roundrobins.cfm. Accessed October 24, 2009.
13. Jennings P, Aydin S, Bennett J, et al. Inter-laboratory comparison of human renal proximal tubule (HK-2) transcriptome alterations due to Cyclosporine A exposure and medium exhaustion. Toxicol In Vitro 2009; 23:486–499.
14. Information Technology Laboratory. Biochange 2008 pilot. Available online at: http://www.itl.nist.gov/iad/894.05/biochange2008/Biochange2008-webpage.htm. Accessed October 24, 2009.
15. Hsu J. Multiple comparisons: theory and methods. London, UK: Chapman & Hall, 1996.
16. Chow SC, Shao J, Wang H. Sample size calculations in clinical research. New York: Marcel Dekker, 2003.
17. Department of Health and Human Services. Available online at: http://grants.nih.gov/grants/guide/pa-files/PAR-08-225.html. Accessed October 24, 2009.