# Legacy Biomarker Qualification Project Status Update[1]

## Administrative Information

*Requesting Organizations*
*Name*: **Biomarkers Consortium, Foundation for the National Institutes of Health (FNIH)** &
**The Radiologic Society of North America, Quantitative Imaging Biomarkers Alliance (RSNA-QIBA)**

*Addresses:* 11400 Rockville Pike, Suite 600
North Bethesda, MD 20852

820 Jorie Blvd
Oak Brook, IL 60523-2251

*Phones:* (301) 402-5311, (630)-571-2670

*Emails*: **foundation@fnih.org** qiba@rsna.org

*Websites*: **https://fnih.org/what-we-do/biomarkers-consortium** **http://www.rsna.org/qiba**

*Primary Contact*
*Name*: **Linda Doody, PhD; Executive Director, Senior Director Clinical Research and Regulatory Affairs**
*Address*: **CCS Associates; 2001 Gateway Place, Ste 350W, San Jose CA, 95110**
*Phone*: **650-691-4400, ext 107**
*Email*: **ldoody@ccsainc.com**

*Alternate Contact*
*Name*: **Ying Tang, PhD; Senior Scientist, Scientific Affairs**
*Address*: **CCS Associates; 2001 Gateway Place, Ste 350W, San Jose CA, 95110**
*Phone*: **650-691-4400, ext 134**
*Email:* **ytang@ccsainc.com**

Submission Date (MM/DD/YYYY): **10/02/2018**

[1] The content you provide in this completed Status Update will be publicly posted as part of the section 507 transparency provisions.

# I. Context of Use

## A. Biomarker Category

*pharmacodynamic/response*

## *B.* Intended Use in Drug Development

*As a primary endpoint for evaluating treatment efficacy/response.*

## C. Context of Use Statement

*Radiologic measurements of whole tumor volume are more precise (reproducible) than unidimensional measurements of tumor diameter. Therefore, longitudinal or serial changes in whole tumor volume during therapy can identify response earlier than corresponding unidimensional measurements, resulting in smaller, more efficient clinical trials. Tumor response or progression as determined by tumor volume can serve as the primary endpoint in well-controlled phase 2 and 3 efficacy studies of cytotoxic, targeted, or immunotherapeutic agents in clinical trials of solid tumors.*

# II. Drug Development Need

*CT imaging technology has significantly improved over the past decades (1). The benefits of imaging for diagnosis, staging, and re-staging cancer are now well established (2, 3). While clinical outcomes remain the gold standard for assessing the value of new treatments, clinical outcomes as an endpoint may not be feasible or optimal in some circumstances. For example, in certain cancer types with a long natural history, it may take years to reach clinical outcome and requires a large number of patients in the studies. Additionally the relationship of clinical outcome to the experimental therapy can be confounded by subsequent therapies, making it difficult to interpret true therapeutic effects.*

*Alternatively, imaging approaches, both qualitative impressions and quantitative analysis, have been proposed to assess the serial changes in tumor burden as an indicator of response to treatment. The current standard method to measure tumor response to therapy using computed tomography (CT) remains Response Evaluation Criteria in Solid Tumors (RECIST), which is based on unidimensional, linear measurements of tumor diameter in the axial plane (4). Because only unidimensional linear measurements are assessed with RECIST, the much higher resolution data offered by modern CT scanners or the advanced image segmentation and visualization methods that can be used on these CT image data sets, available on many commercial workstations, are not fully used (5). The rationale for volumetric approaches to assessing serial changes in tumor burden is multi-factorial. First, most cancers may grow and regress irregularly in three dimensions. Measurements obtained in the axial plane fail to account for growth or regression in the longitudinal axis, whereas volumetric measurements incorporate changes in all dimensions.*

*Secondly, changes in volume are less subject to either reader error or interscan variations. For example, partial response (PR) using RECIST requires a greater than 30% decrease in tumor diameter, which corresponds to 65% reduction in volume of tumor. If one assumes a 21 mm diameter lesion (of 4850 $mm^3$ volume), PR would require that the tumor shrink to a diameter of less than 15 mm, which would correspond to a decrease in volume all the way down to 1770 $mm^3$. The much greater magnitude of volumetric changes is less prone to measurement error than changes in diameter, particularly if the lesions are irregularly shaped or spiculated. As a result of the increased sensitivity and reproducibility, volumetry is likely to be more suited than unidimensional measurements to identify early changes in patients undergoing treatment. Another limitation of RECIST criteria beyond measurement precision is that it was designed for the study of cytotoxic chemotherapies, before the realization of therapeutic success in targeted and immunotherapeutic drugs and the unique radiographic features associated with these drug classes.*

*Studies have shown that it is technically feasible to achieve less than 1% intra- and inter-rater variability when analyzing well-demarcated tumors with simple geometric shapes on a single image set (6). Results from "coffee break" test-retest studies have demonstrated high agreement in volume measurements for pairs of images within subjects acquired after very short time intervals, with 95% limits of relative measurement difference ranging from −12.1% to 13.4%, and a mean relative difference of 0.7% (7). A more recent report of test-retest study in lung cancer patients of the ACRIN 6678 trial using low-dose CT concurred with the previous findings and showed a mean relative volume difference of −0.4% ± 10.5% (mean ± SD), with 95% upper and lower relative measurement difference limits of −21.0% and 20.3% (8). These findings suggest that CT volumetry represents a sufficiently reproducible method in determining tumor burden, and therefore can be relied upon as the basis for imaging biomarkers for predicting therapeutic effects in phase 2 and phase 3 trials.*

*Since tumor volume change represents a downstream event of drug effects independent of drug classes, tumor volume-based biomarkers , if qualified, are likely suitable to be broadly applied in drug development programs of a wide range of solid tumor diseases and drug classes including cytotoxic, targeted, and immunotherapeutic agents in phase 2 and 3 oncology clinical trials. A more sensitive and precise tumor volume-based biomarker has the potential to require fewer subjects for drug efficacy demonstration and to detect drug effects earlier, resulting in smaller and shorter phase 2 or phase 3 trials to reduce the cost and to accelerate oncologic drug development.*

## III. Biomarker Information

### A. Biomarker Name, Source, Type and Description

*Tumor Volume Change as an Imaging Biomarker Predicting Response to Cancer Therapy for Patient Management and Oncologic Drug Development; DDT # (DDTBMQ000011)*

| **Type of Biomarker** (Check relevant type(s)) | | | |
|---|---|---|---|
| | Molecular | × | Radiologic/Imaging |
| | Histologic | | Physiologic Characteristic |
| | Other (please describe): | | |

### B. For molecular biomarkers, please provide a unique ID.

*Not applicable*

### C. Rationale for Biomarker

*Tumor volume measured by CT reflects tumor burden and can be measured quantitatively and with high reproducibility over time.*

*Mechanistic rationale or biologic plausibility for the biomarker*

*Endpoints based on radiographic assessment of the change in tumor burden, such as objective response rate (ORR) and time-to-disease progression (TTP), are frequently used in phase 2 trials to screen for activity of anti-cancer agents. These endpoints have been proven to reasonably predict future clinical outcomes in phase 3 trials in a range of solid cancer types, including colorectal cancer, non-small-cell lung cancer, breast cancer, ovarian cancer and other cancers, for both cytotoxic and targeted therapies (9-12). More recently in the immunotherapeutic setting, immunotherapeutic drugs that conferred ORR advantages in early phase studies went on to show prolonged survival at least in lung cancer and melanoma in phase 3 trials (13-15). These findings suggest that there is a plausible link between tumor burden change and clinical outcomes.*

*Tumor burden has historically been approximated by unidimensional or bidimensional measurement on CT scans to determine response to treatment or disease progression since volumes could not be easily or accurately measured. Technological advances in signal processing and the engineering of multidetector computed tomography (MDCT) devices have resulted in the ability to rapidly acquire high-resolution images, resulting in precise volumetric scanning of anatomic regions. Volumetry is likely to be a substantially more sensitive technique for detecting serial changes in tumor masses than reliance on measurements of lines representing tumor diameters as defined by RECIST. As a result, volumetry may allow earlier and more accurate assessment of clinical outcomes compared with unidimensional measurements used in RECIST.*

*Volumetry could also benefit patients who need alternative treatments when their diseases stop responding to their current regimens.*

*Natural history of the disease indication and associated risk factors*
*We anticipate that the change in tumor burden as measured by CT volumetry can be a pharmacodynamic/response marker in a phase 2 and phase 3 clinical trials in a range of solid tumor diseases, including lung, melanoma, and colorectal and renal cancers treated by cytotoxic, targeted and immunotherapeutic drugs.*

*The magnitude and duration of change in the biomarker required to demonstrate a clinically meaningful effect/impact or outcome.*
*This proposal is to generate evidence to establish the magnitude and duration of change in the biomarker required to demonstrate a clinically meaningful effect/impact on outcome. Our initial hypothesis was to validate the volumetric equivalent of the uni-dimensional RECIST response categories (eg. PD= 20% ↑1D vs 73% ↑3D vs PR=30% 1D ↓ vs 66% ↓3D) as the default threshold assumptions for anatomical volume change in response to a broad range of therapies where tumor shrinkage is the anticipated mechanism of action providing clinical benefit. Secondarily we intended to examine alternative and optimal thresholds (cutpoints) for response or progression based on correlation with clinical outcomes in retrospective analysis of randomized controlled trials. The proposed pharmacodynamic/response markers based on both approaches will be developed and validated within the CT-Vol PACT Project (See Section IX C Ongoing Information Collection)*

*More recently we are exploring alternative methods of assessing the performance of both unidimensional and volumetric tumor measurements using continuous variable analysis methodologies, such as tumor growth/regression kinetic modeling, as potentially superior methods of predicting outcome correlations compared to categorical imaging assessments.*

*Is there an established "baseline" for the biomarker in the target patient population compared to healthy controls? Does another measure of disease progression track with changes in the biomarker that are larger than the standard error in the longitudinal measures? Are baseline measures different from baseline, i.e., is there a clinically validated cut-off or threshold for change in the biomarker? If no, can the results of a patient cohort study be used to develop a statistical model to establish cut-points or a threshold that may be clinically meaningful (see also Section X. Knowledge Gaps in Biomarker Development).*

*The separation of the tumor from its surrounding anatomic structures is made possible by differential radiodensities. In primary NSCLC lesions, Hounsfield units (HU) of the tumor (≥20 HU in general) readily distinguish it from the airspace (−1000 HU) and lung parenchyma (−600 to −700 HU). In metastatic disease, differentiation of the lesion boundaries and volume within an organ of comparable contrast and grey scale value may not be as apparent. Target lesions on CT scans at each patient's baseline study are selected per RECIST, and each lesion is delineated*

*using segmentation software. The longest axial plane diameter (unidimensional measurement) and the volume of a lesion can be automatically calculated by computer programs.*

*Tumor volume can be measured quantitatively with high reproducibility. Using a dataset of 32 NSCLC patients who were scanned twice during a short interval (within 15 minutes) on the same scanner under a presumed no-change condition, the 95% limits of agreements for the computer-aided volumetric measurements on two repeat scans were (−12.1%, 13.4%) (16) by three readers, indicating that changes in tumor volume outside the limits represent true changes. In another study using the same dataset, five readers were instructed to read the scans in a "locked sequential read" manner, i.e., radiologists read the first time point scan, locked their measurements, and then made measurements on the second time point scan while being allowed to review their prior measurements on the first time point scan. Using this workflow, which is more reflective of clinical trials practice, the mean percent difference (± SD) when pooled across both readers (five readers in total) and lesions was 7.4 ± 44.2% (17). A recently published report of a test-retest study in NSCLC cancer patients from the ACRIN 6678 trial using low-dose CT showed a mean relative volume difference of −0.4% ± 10.5% (mean ± SD), with 95% upper and lower relative measurement difference limits of −21.0% and 20.3% (8). The above limits of agreement are substantially narrower than the volumetric equivalent of the uni-dimensional RECIST response categories (eg. PD= 73% ↑3D vs PR= 66% ↓3D), confirming that true tumor volume changes occurred below these thresholds.*

*We are currently in the process of developing a statistical model to establish cut-points or thresholds that are clinically meaningful. We are also developing continuous variable analysis methodologies such as tumor growth/regression kinetic modeling that can predict clinical outcomes (18). We have gathered CT images and clinical outcome data from phase 3 clinical trials sponsored by pharmaceutical companies to support regulatory approval. These include three NSCLC, two colorectal cancer, two renal cell cancer trials of targeted agents; and two melanoma trials of immunotherapeutics, representing over 7,000 patients' images (See Section IX. Evaluation of Existing Biomarker Information: Summaries for details). Of these 10 total trials, two are placebo-controlled studies, while eight use an active comparator to study the response of the investigational drug candidates. We will divide the data into the training set and the validation set. We will use the training dataset to develop statistical models to establish cut-points/thresholds and to develop continuous variable methods that predict clinical outcomes; these cut-points and continuous variables will be confirmed using the validation dataset.*

# IV. Biomarker Measurement Information

## A. General Description of Biomarker Measurement

*Measurement of the tumor volume on CT images should follow the consensus guidelines as described in the QIBA Profile: CT Tumor Volume Change for Advanced Disease (CTV-AD) (19).*

*There are a variety of software packages available that are QIBA complaint, and people using this biomarker could use any of those packages. Examples of the software are described in (20, 21).*

## B. Test/Assay Information

*Indicate whether the biomarker test/assay is one or more of the following:*

i. Laboratory Developed Test (LDT) ○ Yes √No

ii. Research Use Only (RUO) ○ Yes √ No

iii. FDA Cleared/Approved. ○ Yes ○ No √ Don't Know
If yes, provide 510(k)/PMA #:_
Multiple scanner vendors and software providers

iv. If the biomarker is qualified, will the test/assay be performed in a Clinical Laboratory Improvement Amendments (CLIA)–certified laboratory?
○ Yes √ No

v. Is the biomarker test currently under review by the Center for Devices and Radiological Health or the Center for Biologics Evaluation and Research?
○ Yes ○ No √ Don't Know

vi. Is there a standard operating procedure (SOP) for sample collection and storage?
√ Yes ○ No
*Refer to QIBA Profile: CT Tumor Volume Change for Advanced Disease (CTV-AD) should be followed (19)*

vii. Is there a laboratory SOP for the test/assay methodology?
√ Yes ○ No
*Refer to QIBA Profile: CT Tumor Volume Change for Advanced Disease (CTV-AD) should be followed (19)*

## C. Biomarker Measurement

i. Quality Control

*The general procedure described in QIBA Profile: CT Tumor Volume Change for Advanced Disease (CTV-AD) should be followed (19).*

*<u>Precision/reproducibility</u>*

*In a test-retest study using a dataset of 32 NSCLC patients who were scanned twice during a 15 minute interval on the same scanner under a presumed no-change condition, the 95% limits of agreements for the computer-aided volumetric measurements on two repeat scans were (−12.1%, 13.4%) (16). In another study on the same dataset, the mean percent difference (± SD) when pooled across both readers (five readers in total) and lesions was 7.4 ± 44.2% (17). A recently published report of test-retest study in NSCLC cancer patients from the ACRIN 6678 trial using low-dose CT showed a mean relative volume difference of −0.4% ± 10.5%, with upper and lower relative measurement difference limits of −21.0% and 20.3% (8). Other published studies*

*reported results within these ranges (22-28). Also see Section D. Additional Considerations for Radiographic Biomarkers under Performance characteristics including sensitivity, specificity, accuracy and agreement.*

*<u>If cutpoint(s) are used, specify the cutpoint(s) and provide rationale for the cutpoints selected.</u>*

*Our initial hypothesis was to validate the volumetric equivalent of the uni-dimensional RECIST response categories (eg. PD= 20% ↑1D vs 73% ↑3D vs PR=30% 1D ↓ vs 66% ↓3D) as the default threshold assumptions for anatomical volume change. Secondarily we intended to examine alternative and optimal thresholds for response or progression based on correlation with clinical outcomes in retrospective analysis of randomized controlled trials.*

*More recently we are exploring alternative methods of assessing the performance of both unidimensional and volumetric tumor measurements using continuous variable analysis methodologies as potentially superior method of predicting outcome correlations compared to categorical imaging assessments. One example of such biomarkers on the continuous scale is the rate of tumor growth/regression (18).*

ii. Quality Assurance

*<u>Type of test:</u> Tumor volume assessment based on CT imaging is classified as an imaging or radiographic biomarker.*

*<u>SOP</u>: The general procedure described in QIBA Profile: CT Tumor Volume Change for Advanced Disease (CTV-AD) should be followed (19).*

*<u>Detailed description of the specialized software needed (e.g., automated digital image analysis software).</u>*

*Measurement of tumor volume requires three-dimensional segmentation software to separate the tumor from the surrounding anatomic structures and to compute tumor volume. The algorithms that have been evaluated in the QIBA studies ranged from fully automated segmentation algorithms which do not allow any user intervention to semi-automated segmentation algorithms which allow minimal input from the user. Semi-automated segmentation algorithms were further divided into subgroups based on the allowable amount of user input, ranging from those that only allow selection of a seed point(s) for the purpose of initiating segmentation to those that allow various degrees of adjustment to parameters or/and to image boundaries (20, 21). Once the radiologist is satisfied with the contour of the respective tumor, the automated volume assessment tool calculates the volume of the tumor. As an example, the following is an excerpt from the QIBA 1A study (29) briefly describing the workflow. "The 3D volumetric measurements were made using a prototype proprietary semi-automated tool (Oncocare Prototype, Siemens Corporate Research, Princeton, NJ), which included a lesion segmentation component. The 3D measurement process was as follows: the reader (1) defined a*

*seed stroke across the lesion (i.e., a RECIST-like line across the perceived maximum diameter of the lesion), (2) applied the segmentation tools, (3) evaluated the quality of the segmentation, and (4) refined or added seeds strokes and reapplied the segmentation tool until satisfied with the 3D nodule segmentation. The software then provided the estimate of nodule volume."*

*A summary on the 12 tumor volume measurement algorithms in the QIBA 3A study can be found in the Appendix of the publication (20), and is included as an attachment to this document*

iii. Limits, Sources and Quantification of Measurement Error

*The following table (Table 1) is extracted from QIBA Profile: CT Tumor Volume Change for Advanced Disease (CTV-AD) (19). It summarizes major factors that affect volume measurement precision, including the size of tumor, acquisition device, radiologist who performs tumor measurement, and the analysis tool.*

**Table 1 Minimum Detectable Differences for Tumor Volume Changes (Informative)**

| **Tumor Diameter** | **Different Acquisition Device** | | | | Same Acquisition Device | | | |
|---|---|---|---|---|---|---|---|---|
| | Different Radiologist | | **Same Radiologist** | | Different Radiologist | | Same Radiologist | |
| | Different Analysis Tool | Same Analysis Tool | Different Analysis Tool | **Same Analysis Tool** | Different Analysis Tool | Same Analysis Tool | Different Analysis Tool | Same Analysis Tool |
| >50mm | 43% | 24% | 43% | **24%** | 37% | 10% | 37% | 8% |
| 35-49mm | 67% | 33% | 65% | **29%** | 62% | 22% | 60% | 14% |
| 10-34mm | 139% | 120% | 80% | **39%** | 136% | 117% | 75% | 28% |

Notes:

1. Acquisition Device actors being different means the scanner used at the two timepoints were different models (from the same or different vendors). Two scanners with different serial numbers but of the same model are considered to be the same Acquisition Device actor.
2. Precision is expressed here as the repeatability or reproducibility coefficient, depending on the column.
3. A measured change in tumor volume that exceeds the relevant precision value in the table indicates 95% confidence in the presence of a true change.
4. Minimum detectable differences can be calculated from the following formula: $1.96 \times \text{sqrt}(2 \times wCV^2)$, where wCV is estimated from the square root of the sum of the variances from the applicable sources of uncertainty (which makes the assumption that the variance components are additive, an assumption that has not yet been tested).
5. The estimates of the sources of variation were derived from several (QIBA) groundwork studies, some of which were performed on phantoms and some of which were performed on human subjects.

## D. Additional Considerations for Radiographic Biomarkers

Image acquisition, analysis, and interpretation
*For this qualification effort, the CT images are obtained from completed phase 3 trials conducted by pharmaceutical companies to support drug regulatory approval. The acquisition conditions are assumed to meet the industry and regulatory standards.*

*The CT images in DICOM format were segmented by a semi-automated software developed by Drs. Binsheng Zhao and Larry Schwartz (Columbia University); the volume for measurable tumors were calculated and the output is numeric values (16). These tumor volume values are used to study the correlation of longitudinal tumor volume changes with clinical outcomes.*

Assessment of uncertainty including repeatability, reproducibility (e.g., within site, across sites, equipment model/manufacturer) and reader variability.
*See Section IV C i. Biomarker Measurement Information; Biomarker Measurement; Quality Control; and Performance characteristics including sensitivity, specificity, accuracy and agreement in this section below.*

Data to support proposed cutpoint(s) if imaging results are not reported as a continuous variable.
*See III C Biomarker Information; Rationale for Biomarker and IV C i. Biomarker Measurement Information; Biomarker Measurement; Quality Control.*

Performance characteristics including sensitivity, specificity, accuracy and agreement.
*QIBA has organized several studies to quantify the bias and precision of tumor volume measurement using CT scans of either an anthropomorphic thorax phantom or from test-retest studies in lung cancer patients and in colorectal cancer patients. The ACRIN 6678 study sponsored by FNIH also contributes to the understanding of tumor measurement precision in lung cancer. The study results are summarized in Table 2.*

Device imaging performance characteristics such as resolution, field of view, distortion, contrast, depth of penetration, signal to noise ratio and other imaging parameters as necessary.
*The performance characteristics for CT scanner that was used to generate the test-retest "coffee break" dataset of 32 NSCLC patients for reproducibility study was described in (16), and are cited below.*

*"CT scans were obtained with a 16–detector row (LightSpeed 16; GE Healthcare, Milwaukee, Wis) or 64–detector row (VCT; GE Healthcare) scanner, both of which are routinely used at the center. Parameters for the 16–detector row scanner were as follows: tube voltage, 120 kVp; tube current, 299–441 mA; detector configuration, 16 detectors × 1.25-mm section gap; and pitch, 1.375:1. Parameters of the 64–detector row scanner were as follows: tube voltage, 120 kVp; tube current, 298–351 mA; detector configuration, 64 detectors × 0.63-mm section gap; and pitch, 0.984:1. The thoracic images were obtained without intravenous contrast material during a breath hold. Since the second scan was considered as a separate scan, its field of view was set given the patient's second scout image. Adjustment was allowed owing to the patient's position in the scanner. Thin-section (1.25 mm) images were reconstructed with no overlap by using the*

*lung convolution kernel and transferred to our research picture archiving and communication system (PACS) server where Digital Imaging and Communications in Medicine (DICOM) images are stored. These thin-section images were then used for both manual measurement and semiautomated computation of tumor sizes."*

*The performance characteristics for the CT scanners that were used to generate low dose CT images in the multicenter trial (ACRIN protocol 6678; FDG-PET/CT as a Predictive Marker of Tumor Response and Patient Outcome: Prospective Validation in Non-small Cell Lung Cancer) in patients with advanced NSCLC treated with chemotherapy are summarized below. CT volumetric data from this study were analyzed post hoc to produce the reproducibility results reported in (8). The study was being conducted under the well-established policies and procedures of ACRIN for protocol management, site qualification, data management, patient accrual, data and safety monitoring, imaging quality assurance, and evaluation. The site/scanner credentialing and quality control parameters for CT scans used for tumor volumetric measurements are summarized in Table 3.*

Algorithms used to interpret the image or data contained in the image. Please provide a full description of these algorithms and validation data or validation plan to confirm the algorithms function as intended.
Provide the name(s) and version(s) of the software package(s) to be used for image acquisition and analysis

*The semi-automated segmentation software developed by Drs. Binsheng Zhao and Larry Schwartz (Columbia University) will be used to measure tumor volume on CT images for the correlation analysis with clinical outcomes using data from randomized trials (See Section IX. Evaluation of Existing Biomarker Information: Summaries for details). This CT segmentation software has shown to measure tumor volume in lung cancer patients with high reproducibility in a test-retest study; the mean relative difference was of 0.7%, and the 95% limits of agreements on two repeat scans was (−12.1%, 13.4%) (16). This measurement precision is comparable to the nine CT segmentation algorithms evaluated in the QIBA 3A(2) algorithm challenge using the same patient dataset (20). The software was also evaluated along with nine other algorithms from different sources in an algorithm challenge study to measure the volume of synthetic nodule in an anthropomorphic phantom; it performed comparably with other volume calculation algorithms in this setting (21).*

*The software developed by Columbia University Drs. Zhao and Schwartz is described in a publication (16), and algorithms used in the QIBA 3A studies are described in the Appendix of a publication (20). As stated in Section IV. A, there are a variety of software packages available that are QIBA complaint, and people using this biomarker could use any of those packages. Please referred to the two publications (20, 21) for examples of these software packages.*

**Table 2 (A) Technical Performance Validation – Volume Measurement Bias and Precision in Phantom Studies**

| Study Description – Phantom | Summary Results | Status |
|---|---|---|
| **QIBA 1A**: Estimate the bias and variability of volumetric measurements of nodule images by six readers. A total of 40 nodules collected from a single scanner were measured. | Measurements were normalized to a 1D scale for comparison of measurement bias and variance among 1D, 2D and 3D. Relative bias for 3D: −1.8%, −0.4%, −0.7%, −0.4%, and −1.6% for 10-mm spherical, 20-mm spherical, 20-mm elliptical, 10-mm lobulated, and 10-mm spiculated nodules compared to 1.4%, −0.1%, −26.5%, −7.8%, and −39.8% for 1D. The three-dimensional measurements were significantly less biased than 1D for elliptical, lobulated, and spiculated nodules. The relative standard deviations for 3D were 7.5%, 3.9%, 3.6%, 9.7%, and 8.3% compared to 5.7%, 2.6%, 20.3%, 5.3%, and 16.4% for 1D. Unidimensional sizing was significantly less variable than 3D for the lobulated nodule and significantly more variable for the ellipsoid and spiculated nodules. | Completed and published (29) |
| **QIBA 1C**: Estimate the bias and variability of volumetric measurements of images collected from six CT scanners. A total of 462 measurements were made (n=462=6 lesions × [5 scanners × 2 CT protocols+1 scanner × 1 CT protocols] × 7 readers). | The overall percent error for all nodules (n=462) was −6.04±17.60% (mean±SD). The percent error for nodules ≥10 mm (n=308) was –0.59±9.57%, and −16.92±23.89% for nodules <10 mm. Relative bias in pooling the 6 nodules (3 spherical; 3 spiculated) is within a 15% tolerance. On individual nodules, scanner equivalence is found for the larger synthetic lesions (10 mm and 20 mm). Equivalence of the two imaging acquisition protocols supports ACRIN 6678. The study demonstrates in larger lesions (>=10mm diameter) bias and variance can be approximately 15% or less across lesion types, scanners and protocols; it confirms QIBA CT lesion size guidance. | Completed; reported in the BD submission to BQRT 09/30/2012, and as a conference abstract (30) |
| **QIBA 3A(1)–Pilot**: Estimate the bias and variability of volumetric measurements of nodule images collected from a single scanner.<br><br>This is a QIBA organized public challenge. A total of 97 nodules with varying size, shape, and density were measured volumetrically by each of the 12 segmentation algorithms. | This set of images were provided to the participants for the purpose of training their CT segmentation algorithms. The participants were also provided with nodules' volume ground truth values; they were required to record and submit their measured volume results.<br><br>The overall mean percent error (± SD) of volumetric measurements was –1.46 % (±23.94%); the percent error by the individual factors, *i.e.*, nodule size, shape, density, and reconstruction slice thickness, was 0.62% (±21.11%), –3.79% (±21.54%), –1.30% (±21.72%), and –1.34% (±24.02%), respectively, across algorithms. The mean percent errors are below 1% with SDs below 11% when technical conditions satisfy those described in the QIBA CT volumetry Profile. | Completed; reported in the BD submission 09/30/2012. |

| | | |
|---|---|---|
| **QIBA 3A(1)–Pivotal**: Estimate the bias and variability of volumetric measurements of nodule images collected from a single scanner.<br><br>This is the second part of the QIBA organized public challenge. A total of 408 nodules with varying size, shape, density, and reconstruction slice thickness were measured volumetrically by each of the 10 semi-automatic segmentation algorithms with varying degrees of allowable post-segmentation correction. | Ten of the twelve algorithms from the QIBA 3A(1)-Pilot project participated in this Pivotal study.<br><br>The overall mean percent error of volumetric measurements across nodule characteristics (nodule size, shape, density, and reconstruction slice thickness) and algorithms was 1.04% [95% CI (0.06–2.13%)]. When only those nodules that satisfy the QIBA CT profile (density > −630 HU; size ≥ 10 mm; non-irregular shaped; reconstruction slice thickness < 3 mm) were included in the analysis, the overall percent error of volumetric measurements was reduced to –0.65% [95% CI (–1.66, 0.36%)].<br><br>Over all nodules meeting the QIBA Profile, the repeatability coefficient (RC) was 9.0% for two measurements of nodule volume by the same algorithm; the between-algorithm reproducibility coefficient (RDC) was 22.1% for measuring a nodule by different algorithms.<br><br>Algorithm type did not affect bias substantially; however, it was an important factor in measurement precision. Algorithm precision was notably better as tumor size increased, worse for irregularly shaped tumors, and on the average better for type 1 algorithms where post-segmentation correction was not allowed. Over all nodules meeting the QIBA Profile, precision, as measured by the repeatability coefficient, was 9.0% compared to 18.4% overall.<br><br>The study concluded that the results achieved in this study, using a heterogeneous set of measurement algorithms, support QIBA quantitative performance claims in terms of volume measurement repeatability for nodules meeting the QIBA Profile criteria. | Completed; reported in the BD submission 09/30/2012 and also published in (21) |

**Table 2 (B) Technical Performance Validation – Volume Measurement Precision Evaluated with Clinical Patient Data**

| Study Description – Patient Data | Summary Results | Status |
|---|---|---|
| **QIBA 1B**: Estimate the test/retest measurement variability of lesions from 32 non-small cell lung cancer (NSCLC) patients who were scanned twice within 15 minutes ("no change" condition). Five readers measured the volume according to two different reading schemes: (1) random presentation of scans, i.e., independent reads, and (2) locked, sequential read of scans from the same lesion. | This work shows that variability within a sizing method may be influenced by the reading paradigm. The 1D sizing method results do not change significantly or substantially across reading paradigms. The means of percent difference were 2.75% with 95% CI [−2.34%, 7.83%] in the independent reading, and 2.52% with 95% CI [−0.28%, 5.33%] in the locked sequential reading. However, volume measurements do change substantially and differences are lower for the locked sequential reading paradigm, but this did not reach statistical significance (P = .067). In the summary statistic for volume measurements, the means were 23.40% with 95% CI [−2.36%, 52.34%] in the independent reading, and 7.42% with 95% CI [−0.98%, 15.82%] in the locked sequential reading. The bias of measurements in this study cannot be assessed as the true lesion size is unknown.<br><br>It should be noted, unlike the report by Petrick et al. (29) where 3D percent change is normalized to a 1D scale to allow comparison, this study reported percent changes in their original scales. | Completed, and published (17) |
| **QIBA 3A(2)**: Estimate the test/retest measurement variability of lesions from 32 NSCLC patients who were scanned twice within 15 minutes (the same dataset as 1B described above). This study is organized as a public challenge. Intra-algorithm and inter-algorithm variability was analyzed for 12 diverse tumor | The approximate tumor diameters ranged from 8 to 65 mm. Intra-algorithm repeatability ranged from 13% to 24% for nine of the 12 algorithms, with most algorithms demonstrating improved repeatability as the tumor size increased. Change in tumor volume can be measured with confidence to within ±14% using any of these nine algorithms on tumor sizes greater than 10 mm. | Completed, and published (20) |

| | | |
|---|---|---|
| segmentation algorithms from 11 academic and commercial participating members. | | |
| **QIBA 3B**: Inter- and intra-reader variability of volumetric measurement of lesions in lungs, liver and/or lymph nodes in subjects with metastatic colorectal cancer.<br><br>Three readers measured each scan volumetrically for assessment of inter-reader variability; two of the readers repeated measurements for assessment of intra-reader variability. | Using RECIST, three radiologists selected target lesions and measured "uni" (maximal diameter), "bi" (product of maximal diameter and maximal perpendicular diameter), and "vol" (volume) on baseline and 6-week post-therapy scans in the following ways: (i) each radiologist independently selected and measured target lesions and (ii) one radiologist's target lesions were blindly re-measured by the others. Variability in relative change of tumor measurements was analyzed using linear mixed effects models. The model-based estimate for limits of agreement was ±1.96 times the estimate of the within-patient SD, that is, the residual SD.<br><br>Of 198 target lesions total from 29 patients, 33% were selected by all three, 28% by two, and 39% by one radiologist. With independent selection, the variability in relative change of tumor measurements was 11% (uni), 19% (bi), and 22% (vol), respectively. When measuring the same lesions, the corresponding numbers were 8%, 14%, and 12%. | Completed, and published (31) |
| **ACRIN 6678**: test-retest variability of volumetric measurements in advanced NSCLC subjects. The dataset of 34 patients from this study was combined with that of 40 patients from a multicenter Merck MK-0646-008 trail of a comparable cohort. | Repeat scans of 71 primary tumors (1 primary tumor per patient) and 5 additional lesions from low-dose CT images were analyzed. The mean anatomic volume was 52.4 $cm^3$ (median, 37.5 $cm^3$; SD, 53.0 $cm^3$). The repeatability of each metric was assessed with Bland–Altman analysis by reporting the mean and SD of the differences between the two measurements. The anatomic volume determination had a repeatability of −0.4% ± 10.5%, with upper and lower repeatability limits of +20.3% and −21.0%. | Completed, and published (8) |

**Table 3 Site/Scanner Credentialing and Quality Control Parameters for CT Scans used for Tumor Volumetric Measurements**

| DICOM Tag # | Parameter | GE | | | Phillips | | SIEMENS | | | TOSHIBA |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Ultra** 8-slice/ 0.5 sec | **LS 16** 16-slice/ 0.5 sec | **VCT(64)** 64-slice/0.5 sec | **Brilliance** 16-slice/ 0.5 sec 16 x 0.75 | **Brilliance** 64 slice/0.5 sec 16 × 0.75 | **Sensation 16** 16 × 0.75 | **Sensation 40** 40 × 0.6 (beam collimation 20 × 0.6) | **Sensation 64** 64 × 0.6 (beam collimation 32 × 0.6) | **Aquillon** 16-slice/0.5 sec |
| 0018,0050 | Nominal Reconstructed Slice Width[1] | 1–1.5 mm | | | 1–1.5 mm | | 1–1.5 mm | | | 1–1.5 mm |
| | | | | | | | | | | |
| 0020,1041 | Reconstructed Interval[1] | 0–20% overlap | | | 0–20% overlap | | 0–20% overlap | | | 0–20% overlap |
| | | | | | | | | | | |
| 0028,0030 | Voxel Size[1] | 0.55–0.75 mm | | | 0.55–0.75 mm | | 0.55–0.75 mm | | | 0.55–0.75 mm |
| | | | | | | | | | | |
| – | Motion/Breathing Artifact[1] | None | | | None | | None | | | None |
| | | | | | | | | | | |
| – | Intravenous Contrast Media[1] | None | | | None | | None | | | None |
| | | X-ray Tube Current × Exposure Time | | | Exposure | | Exposure | | | X-ray Tube Current × Exposure Time |
| Scanner-dependent | mAs (Regular-Large)[2] | 135–220 | 95–245 | 95–245 | 120–310 | 100–260 | 120–310 | 100–260 | 100–260 | 120–310 |
| | | | | | | | | | | |
| 0010,0000 | KVP[2] | 120 | | | 120 | | 120 | | | 120 |
| | | | | | | | | | | |
| 0010,1210 | Reconstruction Algorithm[2] | STD | | | B | | B30 | | | FC10 |

[1]Violation of values/value ranges disqualifies CT scan series.

[2]Violation of values/value ranges may not disqualify CT scan series (unless violation is excessive). Comment on lower or higher than recommended values (*e.g.,* for KVP of 140 you may comment as to use similar KVP of 140 for the follow up of the same ca but try to follow the protocol for next individuals, *i.e.,* using KVP of 120 for the future cases.)

## VIII. Assessment of Benefits and Risks

### A. Anticipated Benefits

*If the utility of this CT tumor volume-based biomarker can be confirmed as a highly reproducible pharmacodynamic/response marker, it has the potential to facilitate oncologic drug development by shortening phase 2 trials of investigational drugs and detecting clinical benefit earlier in phase 3 investigations, resulting in reduction in clinical trial time and costs.*

*This biomarker can benefit patients with cancer who need to know as soon as possible whether or not they are benefiting from new treatments. It will help patients seek alternatives sooner once their therapeutic regimens become futile.*

### B. Anticipated Risks

*False declaration of treatment response (false positive) by the biomarker may mislead the physician to continue the ineffective treatment, patients to endure unnecessary toxicity and lose the window of opportunity for potential alternative therapy.*

*False declaration of no response (false negative) may result in premature termination of an effective treatment and its associated benefits.*

### C. Risk Mitigation Strategy

*There are steps that can be implemented to reduce false positive and false negative determination of patient response/progression. These include rigorous quality control steps in image collection, using algorithms with high precision in tumor volume measurement, robust statistical methodologies, and high quality of imaging and clinical data being used for biomarker development. These specifications have been documented in the QIBA Profile (19).*

### D. Conclusions

*We anticipate that this biomarker will have higher precision, and be more sensitive and specific than the currently accepted RECIST-based endpoints in predicting phase 2 and phase 3 outcomes in solid tumors. Therefore, the benefit and risk balance is in favor of this biomarker.*

## IX. Evaluation of Existing Biomarker Information: Summaries

### A. Pre-Clinical Information, as appropriate

*Not applicable.*

### B. Completed Clinical Information, as appropriate

*Please refer to Table 2 (B) Technical Performance Validation – Volume Measurement Precision Evaluated with Clinical Patient Data under Section D. Additional Considerations for Radiographic Biomarkers.*

### *C.* Summary of Ongoing Information Collection/Analysis Efforts

*To establish a systematic approach to develop and validate imaging-based biomarkers to improve upon RECIST, the Foundation for the National Institutes of Health (FNIH) Biomarkers Consortium initiated a collaborative research partnership entitled Vol-PACT (Volumetric CT for Precision Analysis of Clinical Trial Results). Vol-PACT is collecting imaging data and associated patient outcomes data from large and completed landmark phase 3 trials in several measurable solid tumors (Table 4). These trials were sponsored by pharmaceutical companies; data of these trials are of regulatory quality and have been reviewed by the FDA. The use of archived data is a cost-effective approach, and eliminates the extensive resources and time needed to conduct prospective trials for purposes of biomarker development and validation. The aim is to retrospectively analyze these high quality data and comprehensively study biomarkers/metrics in the context of unidimensional and volumetric tumor measurements in their ability to predict clinical outcomes.*

*The CT images, which are collected centrally on most trials, are transferred from various imaging core laboratories to an academic laboratory for tumor measurement. Next, images are re-analyzed in a semi-automated fashion with computer-generated contouring to determine unidimensional and volumetric measurements for each lesion at each time point. These imaging measurement readouts are used to study the correlation of the proposed biomarkers with clinical outcomes in order to develop pharmacodynamics/response biomarkers.*

*We have obtained access to both DICOM images and clinical metadata for three lung cancer trials (Lux-Lung 1, Lux-Lung 3, Lux-Lung 6), three colorectal cancer trials (VELOUR, PRIME, 20020408), two renal cell cancer trials (VEG105192, COMPARZ), and two melanoma immunotherapy trials (Keynote 002, Keynote 006)(Table 4).*

## X. Knowledge Gaps in Biomarker Development

### A. List and describe any knowledge gaps, including any assumptions, that exist in the application of the biomarker for the proposed COU

*There is a strong rationale that the change in tumor burden reflects the disease status, therefore it is plausible that a biomarker based on the change in tumor burden can predict for patient's response to cancer treatment. We have systematically studied the reproducibility of tumor volume measurement and understand that ultimate evidence to support the COU requires correlation of the biomarker with clinical outcomes. We have collected imaging and clinical*

*metadata (Table 4), and are currently conducting retrospective analysis to validate the biomarkers for the proposed COU.*

**B. List and describe the approach/tools you propose to use to fill in the above-named gaps when evidence is unknown or uncertain,** (i.e., statistical measures and models, meta-analysis from other clinical trials).

*See Section IX C. Summary of Ongoing Information Collection/Analysis Efforts.*

**C. Describe the status of other work currently underway and planned for the future toward qualification of this biomarker for the proposed context of use.**

*See Section IX C. Summary of Ongoing Information Collection/Analysis Efforts***.**

*We have obtained access to imaging data and associated patient outcomes data from large and completed landmark phase 3 trials in several measurable solid tumors (Table 4). To our knowledge, this is the largest collection of images and associated clinical data thus far for the purpose of developing imaging-based biomarkers for drug development. Since the tumor burden change is an indicator of disease status, we anticipate that this biomarker will have general application in drug development programs in a range of solid tumor diseases, including lung cancer, colorectal cancer, renal cell carcinoma, melanoma, and other cancer types.*

## Table 4 Clinical Trial Data Collection

| Trial Sponsor | Disease | Drug | Trial ID | N | Primary Endpoint | OS HR (95% CI) | PFS HR (95% CI) | Data Analysis | Publication |
|---|---|---|---|---|---|---|---|---|---|
| Sanofi | CRC | FOLFIRI +/- aflibercept | VELOUR | 1226 | OS | 0.817 (0.714, 0.935), p=0.0032 | 0.758 (0.661, 0.869), p=0.0007 | data analysis underway | (32) |
| GSK /Novartis | RCC | Pazopanib vs. placebo | VEG105192 | 435 | PFS | 0.91 (0.71, 1.16) | 0.46 (0.34, 0.62), p<0.0001 | data analysis underway | (33, 34) |
| GSK /Novartis | RCC | Pazopanib vs. sunitinib | COMPARZ | 1110 | PFS | 0.91 (0.76, 1.08) | 1.05 ( 0.90, 1.22) | data analysis underway | (35) |
| Amgen | CRC | FOLFOX +/- panitumumab | PRIME | 1183 | PFS | 0.83 (0.67, 1.02), p=0.072 | 0.80 (0.66 to 0.97), p=0.02 | data analysis underway | (36) |
| Amgen | CRC | BSC+/- panitumumab | 20020408 | 463 | PFS | 1.00 (0.82 to 1.22) | 0.54 (0.44, 0.66), p<0.0001 | data analysis underway | (37) |
| BI | NSCLC | Afatinib vs. placebo | Lux-Lung1 | 585 | OS | 1.08 (0.86, 1.35), p=0.74 | 0.38 (0.31-0.48), p<0.0001 | data analysis underway | (38) |
| BI | NSCLC | Afatinib vs. pemetrexed + cisplatin | Lux-Lung3 | 345 | PFS | 0.88 (0.66, 1.17), p=0.39 | 0.58 (0.43, 0.78), p=0.001 | data analysis underway | (39, 40) |
| BI | NSCLC | Afatinib vs. gemcitabine + cisplatin | Lux-Lung6 | 364 | PFS | 0.93 (0.72, 1.22), p=0.61 | 0.28 (0.2, 0.39), p<0.0001 | data analysis underway | (40, 41) |
| Merck | Mel | Pembrolizumab vs. Ipilimumab | Keynote 006 | 834 | OS, PFS | 0.68 ( 0.53, 0.87), p=0.0009; 0.68 (0.53, 0.86), p=0.0008 | 0.61 (0.50, 0.75), p<0.0001; 0.61 (0.50, 0.75), p<0.0001 | data analysis underway | (42) |
| Merck | Mel | Pembrolizumab vs. chemo | Keynote 002 | 540 | OS, PFS | 0.86 (0.67, 1.10), p=0.117; 0.74 (0.57, 0.96), p=0.011 | 0.58 (0.46, 0.73), p<0.0001; 0.47 (0.37, 0.60), p<0.0001 | data analysis underway | (43) |

**APPENDIX I.**

## List of publications


1. Fuchs T, Kachelriess M, Kalender WA. Technical advances in multi-slice spiral CT. Eur J Radiol. 2000;36(2):69-73.
2. Tempany CM, McNeil BJ. Advances in biomedical imaging. Jama. 2001;285(5):562-7.
3. Diederich S, Wormanns D. Impact of low-dose CT on lung cancer screening. Lung Cancer. 2004;45 Suppl 2:S13-9.
4. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228-47.
5. McNitt-Gray MF, Bidaut LM, Armato SG, Meyer CR, Gavrielides MA, Fenimore C, et al. Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error. Transl Oncol. 2009;2(4):216-22.
6. Goodman LR, Gulsun M, Washington L, Nagy PG, Piacsek KL. Inherent variability of CT lung nodule measurements in vivo using semiautomated volumetric measurements. AJR Am J Roentgenol. 2006;186(4):989-94.
7. Zhao B, Schwartz LH, Larson SM. Imaging surrogates of tumor response to therapy: anatomic and functional biomarkers. J Nucl Med. 2009;50(2):239-49.
8. Desseroit MC, Tixier F, Weber WA, Siegel BA, Cheze Le Rest C, Visvikis D, et al. Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non-Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort. J Nucl Med. 2017;58(3):406-11.
9. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-Analysis Group in Cancer. Lancet. 2000;356(9227):373-8.
10. Goffin J, Baral S, Tu D, Nomikos D, Seymour L. Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. Clin Cancer Res. 2005;11(16):5928-34.
11. Paesmans M, Sculier JP, Libert P, Bureau G, Dabouis G, Thiriaux J, et al. Response to chemotherapy has predictive value for further survival of patients with advanced non-small cell lung cancer: 10 years experience of the European Lung Cancer Working Party. Eur J Cancer. 1997;33(14):2326-32.
12. El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. J Clin Oncol. 2008;26(8):1346-54.
13. Robert C, Ribas A, Wolchok JD, Hodi FS, Hamid O, Kefford R, et al. Anti-programmed-death-receptor-1 treatment with pembrolizumab in ipilimumab-refractory advanced melanoma: a randomised dose-comparison cohort of a phase 1 trial. Lancet. 2014;384(9948):1109-17.
14. Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. N Engl J Med. 2015;372(26):2521-32.
15. Herbst RS, Baas P, Kim DW, Felip E, Perez-Gracia JL, Han JY, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet. 2016;387(10027):1540-50.

16. Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology. 2009;252(1):263-72.
17. McNitt-Gray MF, Kim GH, Zhao B, Schwartz LH, Clunie D, Cohen K, et al. Determining the Variability of Lesion Size Measurements from CT Patient Data Sets Acquired under "No Change" Conditions. Transl Oncol. 2015;8(1):55-64.
18. Wilkerson J, Abdallah K, Hugh-Jones C, Curt G, Rothenberg M, Simantov R, et al. Estimation of tumour regression and growth rates during treatment in patients with advanced prostate cancer: a retrospective analysis. Lancet Oncol. 2017;18(1):143-54.
19. CT Volumetry Technical Committee. CT Tumor Volume Change Profile - 2018, Technically Confirmed Profile. Quantitative Imaging Biomarkers Alliance, June 22, 2018. Available at: http://qibawiki.rsna.org/index.php/Profiles.
20. Buckler AJ, Danagoulian J, Johnson K, Peskin A, Gavrielides MA, Petrick N, et al. Inter-Method Performance Study of Tumor Volumetry Assessment on Computed Tomography Test-Retest Data. Acad Radiol. 2015;22(11):1393-408.
21. Athelogou M, Kim HJ, Dima A, Obuchowski N, Peskin A, Gavrielides MA, et al. Algorithm Variability in the Estimation of Lung Nodule Volume From Phantom CT Scans: Results of the QIBA 3A Public Challenge. Acad Radiol. 2016;23(8):940-52.
22. Marten K, Auer F, Schmidt S, Kohl G, Rummeny EJ, Engelke C. Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria. Eur Radiol. 2006;16(4):781-90.
23. Oda S, Awai K, Murao K, Ozawa A, Yanaga Y, Kawanaka K, et al. Computer-aided volumetry of pulmonary nodules exhibiting ground-glass opacity at MDCT. AJR Am J Roentgenol. 2010;194(2):398-406.
24. Gietema HA, Schaefer-Prokop CM, Mali WP, Groenewegen G, Prokop M. Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection CT-- influence of inspiration level, nodule size, and segmentation performance. Radiology. 2007;245(3):888-94.
25. Hein PA, Romano VC, Rogalla P, Klessen C, Lembcke A, Dicken V, et al. Linear and volume measurements of pulmonary nodules at different CT dose levels - intrascan and interscan analysis. Rofo. 2009;181(1):24-31.
26. Wang Y, van Klaveren RJ, van der Zaag-Loonen HJ, de Bock GH, Gietema HA, Xu DM, et al. Effect of nodule characteristics on variability of semiautomated volume measurements in pulmonary nodules detected in a lung cancer screening program. Radiology. 2008;248(2):625-31.
27. Revel MP, Lefort C, Bissery A, Bienvenu M, Aycard L, Chatellier G, et al. Pulmonary nodules: preliminary experience with three-dimensional evaluation. Radiology. 2004;231(2):459-66.
28. Nishino M, Guo M, Jackman DM, DiPiro PJ, Yap JT, Ho TK, et al. CT tumor volume measurement in advanced non-small-cell lung cancer: Performance characteristics of an emerging clinical tool. Acad Radiol. 2011;18(1):54-62.
29. Petrick N, Kim HJ, Clunie D, Borradaile K, Ford R, Zeng R, et al. Comparison of 1D, 2D, and 3D nodule sizing methods by radiologists for spherical and complex nodules on thoracic CT phantom images. Acad Radiol. 2014;21(1):30-40.
30. Fenimore C, McNitt-Gray MF, Lu J, Kim G, Clunie DA, Gavrielides MA, et al. Clinician Sizing of Synthetic Nodules to Evaluate CT Interscanner Effects. RSNA 2012.abstr SSG15-01. Available at: http://archive.rsna.org/2012/12035113.html.
31. Zhao B, Lee SM, Lee HJ, Tan Y, Qi J, Persigehl T, et al. Variability in assessing treatment response: metastatic colorectal cancer as a paradigm. Clin Cancer Res. 2014;20(13):3560-8.

32. Van Cutsem E, Tabernero J, Lakomy R, Prenen H, Prausova J, Macarulla T, et al. Addition of aflibercept to fluorouracil, leucovorin, and irinotecan improves survival in a phase III randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. J Clin Oncol. 2012;30(28):3499-506.
33. Sternberg CN, Davis ID, Mardiak J, Szczylik C, Lee E, Wagstaff J, et al. Pazopanib in locally advanced or metastatic renal cell carcinoma: results of a randomized phase III trial. J Clin Oncol. 2010;28(6):1061-8.
34. Sternberg CN, Hawkins RE, Wagstaff J, Salman P, Mardiak J, Barrios CH, et al. A randomised, double-blind phase III study of pazopanib in patients with advanced and/or metastatic renal cell carcinoma: final overall survival results and safety update. Eur J Cancer. 2013;49(6):1287-96.
35. Motzer RJ, Hutson TE, Cella D, Reeves J, Hawkins R, Guo J, et al. Pazopanib versus sunitinib in metastatic renal-cell carcinoma. N Engl J Med. 2013;369(8):722-31.
36. Douillard JY, Siena S, Cassidy J, Tabernero J, Burkes R, Barugel M, et al. Randomized, phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: the PRIME study. J Clin Oncol. 2010;28(31):4697-705.
37. Van Cutsem E, Peeters M, Siena S, Humblet Y, Hendlisz A, Neyns B, et al. Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. J Clin Oncol. 2007;25(13):1658-64.
38. Miller VA, Hirsh V, Cadranel J, Chen YM, Park K, Kim SW, et al. Afatinib versus placebo for patients with advanced, metastatic non-small-cell lung cancer after failure of erlotinib, gefitinib, or both, and one or two lines of chemotherapy (LUX-Lung 1): a phase 2b/3 randomised trial. Lancet Oncol. 2012;13(5):528-38.
39. Sequist LV, Yang JC, Yamamoto N, O'Byrne K, Hirsh V, Mok T, et al. Phase III study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with EGFR mutations. J Clin Oncol. 2013;31(27):3327-34.
40. Yang JC, Wu YL, Schuler M, Sebastian M, Popat S, Yamamoto N, et al. Afatinib versus cisplatin-based chemotherapy for EGFR mutation-positive lung adenocarcinoma (LUX-Lung 3 and LUX-Lung 6): analysis of overall survival data from two randomised, phase 3 trials. Lancet Oncol. 2015;16(2):141-51.
41. Wu YL, Zhou C, Hu CP, Feng J, Lu S, Huang Y, et al. Afatinib versus cisplatin plus gemcitabine for first-line treatment of Asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-Lung 6): an open-label, randomised phase 3 trial. Lancet Oncol. 2014;15(2):213-22.
42. Schachter J, Ribas A, Long GV, Arance A, Grob JJ, Mortier L, et al. Pembrolizumab versus ipilimumab for advanced melanoma: final overall survival results of a multicentre, randomised, open-label phase 3 study (KEYNOTE-006). Lancet. 2017;390(10105):1853-62.
43. Hamid O, Puzanov I, Dummer R, Schachter J, Daud A, Schadendorf D, et al. Final analysis of a randomised trial comparing pembrolizumab versus investigator-choice chemotherapy for ipilimumab-refractory advanced melanoma. Eur J Cancer. 2017;86:37-45.

## Attachment(s)

Buckler AJ, Danagoulian J, Johnson K, Peskin A, Gavrielides MA, Petrick N, et al. Inter-Method Performance Study of Tumor Volumetry Assessment on Computed Tomography Test-Retest Data. Acad Radiol. 2015;22(11):1393-408

# Inter-Method Performance Study of Tumor Volumetry Assessment on Computed Tomography Test-Retest Data

Andrew J. Buckler, MS, Jovanna Danagoulian, PhD, Kjell Johnson, PhD, Adele Peskin, PhD, Marios A. Gavrielides, PhD, Nicholas Petrick, PhD, Nancy A. Obuchowski, PhD, Hubert Beaumont, PhD, Lubomir Hadjiiski, PhD, Rudresh Jarecha, DNB, DMRE, Jan-Martin Kuhnigk, PhD, Ninad Mantri, MS, Michael McNitt-Gray, PhD, Jan H. Moltz, PhD, Gergely Nyiri, MS, Sam Peterson, MS, Pierre Tervé, MS, Christian Tietjen, PhD, Etienne von Lavante, PhD, Xiaonan Ma, MS, Samantha St. Pierre, BS, Maria Athelogou, PhD

**Rationale and objectives:** Tumor volume change has potential as a biomarker for diagnosis, therapy planning, and treatment response. Precision was evaluated and compared among semiautomated lung tumor volume measurement algorithms from clinical thoracic computed tomography data sets. The results inform approaches and testing requirements for establishing conformance with the Quantitative Imaging Biomarker Alliance (QIBA) Computed Tomography Volumetry Profile.

**Materials and Methods:** Industry and academic groups participated in a challenge study. Intra-algorithm repeatability and inter-algorithm reproducibility were estimated. Relative magnitudes of various sources of variability were estimated using a linear mixed effects model. Segmentation boundaries were compared to provide a basis on which to optimize algorithm performance for developers.

**Results:** Intra-algorithm repeatability ranged from 13% (best performing) to 100% (least performing), with most algorithms demonstrating improved repeatability as the tumor size increased. Inter-algorithm reproducibility was determined in three partitions and was found to be 58% for the four best performing groups, 70% for the set of groups meeting repeatability requirements, and 84% when all groups but the least performer were included. The best performing partition performed markedly better on tumors with equivalent diameters greater than 40 mm. Larger tumors benefitted by human editing but smaller tumors did not. One-fifth to one-half of the total variability came from sources independent of the algorithms. Segmentation boundaries differed substantially, not ony in overall volume but also in detail.

**Conclusions:** Nine of the 12 participating algorithms pass precision requirements similar to what is indicated in the QIBA Profile, with the caveat that the present study was not designed to explicitly evaluate algorithm profile conformance. Change in tumor volume can be measured with confidence to within $\pm$14% using any of these nine algorithms on tumor sizes greater than 10 mm. No partition of the algorithms was able to meet the QIBA requirements for interchangeability down to 10 mm, although the partition comprising best performing algorithms did meet this requirement for a tumor size of greater than approximately 40 mm.

**Key Words:** CT; volumetry; lung cancer; quantitative imaging; segmentation.



**Acad Radiol 2015;** 22:1393–1408

From the Elucid Bioimaging Inc., 225 Main Street, Wenham, MA 01984 (A.J.B., J.D., X.M., S.S.P.); Arbor Analytics LLC, Ann Arbor, Michigan (K.J.); National Institute of Standards and Technology, Boulder, Colorado (A.P.); U.S. Food and Drug Administration, Silver Spring, Maryland (M.A.G., N.P.); Cleveland Clinic, Cleveland, Ohio (N.A.O.); MEDIAN Technologies, Valbonne, France (H.B.); Department of Radiology, University of Michigan, Ann Arbor, Michigan (L.H.); Perceptive Informatics, Sundew Properties SEZ Pvt Ltd Mindspace, Hyderabad, Andhra Pradesh, India (R.J.); Fraunhofer MEVIS, Institute for Medical Image Computing, Bremen, Germany (J.-M.K., J.H.M.); ICON Medical Imaging, Warrington, Pennsylvania (N.M.); Department of Radiology, University of California at Los Angeles, Los Angeles, California (M.M.-G.); GE Healthcare, Buc, France (G.N.); Vital Images, Inc., Los Angeles, California (S.P.); KEOSYS, Saint-Herblain, France (P.T.); Siemens AG, Healthcare Sector, Imaging and Therapy Division, Forchheim, Germany (C.T.); Mirada Medical Ltd., Oxford Center for Innovation, Oxford, United Kingdom (E.v.L); and Definiens AG, München, Germany (M.A.). Received March 2, 2015; accepted August 7, 2015. Disclaimer: Certain commercial equipment, instruments, materials, or software are identified in this article to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, Food and Drug Administration, or any other coauthor nor does it imply that the materials or equipment identified are necessarily the best available for the purpose. **Address correspondence to:** A.J.B. e-mail: andrew.buckler@elucidbio.com


http://dx.doi.org/10.1016/j.acra.2015.08.007

Lung tumor volume change assessed with computed tomography (CT) has potential as a quantitative imaging biomarker to improve diagnosis, therapy planning, and monitoring of treatment response (1,2). Tumor volume change as a predictor of outcome has been of interest for some time (3–5).

To establish confidence in algorithmic analysis for CT volumetry as a rigorously defined assay useful for clinical and research purposes, volume measurement algorithms need to be characterized in terms of both bias and variability. Measurement error on serial CT scans can be affected by a number of interrelated factors, including imaging parameters, tumor characteristics, and/or measurement procedures (6–8). These effects must be understood and quantified. A number of technical studies have been performed toward this goal (9–32).

The Quantitative Imaging Biomarker Alliance (QIBA) (33) has defined standard procedures for reliably measuring lung tumor volume changes in a document called a profile. The CT volumetry profile is based in part on the available literature and on the "groundwork" studies conducted by QIBA itself (34). Groundwork studies of algorithm performance organized as public challenges have been conducted under the moniker of "3A." The first 3A study was conducted to estimate intra-algorithm and inter-algorithm bias and variability using phantom data sets (Athelogou, PhD, manuscript under review, 2015). Algorithms used by participating groups were applied to CT scans of synthetic lung tumors in anthropomorphic phantoms. Although such a study design was effective for estimating bias because ground truth was known, phantom studies are likely to underestimate the biological variability typically seen in clinical data sets. More recently, QIBA has undertaken studies on the analysis of clinical data. The QIBA "1B" study was undertaken to compare two reading paradigms, independent readings at both time points versus locked sequential readings, using a test-retest design (35). Readers in the QIBA 1B study used a single algorithm. The present study, known as the "second" 3A, combines the algorithm performance challenge approach established by the first 3A study using the same clinical data as were used in 1B. The goal of the present study was to quantify the error when a tumor with no biological change in size was imaged twice and each image was measured by the same or multiple algorithms.

Intra-algorithm and inter-algorithm variability was analyzed using data from 12 diverse tumor segmentation algorithms from 12 academic and commercial participating groups for measuring volume. The algorithms included semiautomated algorithms with and without postsegmentation manual correction. The analysis of algorithm performance conducted in this study complements the other groundwork studies in establishing performance claims for the QIBA Profile.

In the following section, we describe the statistical methods and open-source informatics tool used to conduct the study as a challenge problem. The estimated intra-algorithm repeatability and inter-algorithm reproducibility are presented in Results section, which also describes a comparison of the segmentation boundaries themselves for the subset of algorithms where tumor segmentations were submitted.

## MATERIALS AND METHODS

### *Data collection*

Thirty-one subjects with non-small cell lung cancer were evaluated in a test-retest design. The cases were contributed to the Reference Image Database to Evaluate Therapy Response (RIDER) database from Memorial Sloan Kettering Cancer Center, acquired in a previously conducted study (36). Each patient was scanned twice within a short period of time (<15 minutes) on the same scanner and the image data were reconstructed with thin sections (<1.5 mm). Because the time interval between repeat scans is small, the actual volume of the tumor is the same in each scan (a zero-change scenario).

CT scans were obtained with a 16-detector row (LightSpeed 16; GE Healthcare, Milwaukee, Wisconsin) or 64-detector row (VCT; GE Healthcare) scanner. Parameters for the 16-detector row scanner were as follows: peak voltage across the x-ray tube, 120 kVp; tube current, 299–441 mA; detector configuration, 16 detectors × 1.25-mm section gap; and pitch, 1.375. Parameters for the 64-detector row scanner were as follows: tube voltage, 120 kVp; tube current, 298–351 mA; detector configuration, 64 detectors × 0.63-mm section gap; and pitch, 0.984. The thoracic images were obtained without intravenous contrast material during a breath hold. Because the second scan was considered as a separate scan, its field of view was set given the patient's second scout image. Adjustment was allowed owing to the patient's position in the scanner. Thin-section (1.25 mm) images were reconstructed with no overlap by using filtered back projection with the lung convolution kernel and transferred to the research picture archiving and communication system server where digital imaging and communications in medicine images were stored.

One tumor per subject was selected for measurement by the clinical staff at Memorial Sloan Kettering. Among them, most were primary lung cancers but three were metastatic tumors (used because the primary tumors were nonmeasureable, as defined by the Response Evaluation Criteria in Solid Tumors criteria). The data set includes tumors that are distinct and solitary as well as others with attachment to various structures including bronchus, chest wall, and mediastinum. The approximate tumor diameters ranged from 8 to 65 mm, as calculated by the equivalent diameter were a sphere to include the same volume.

The shapes of the selected tumors ranged from simple and isolated to complex and cavitated. To facilitate comparison of results to the prior QIBA 1B study, the tumors were further subdivided according to whether they met the following "measurability" criteria defined in the profile:

tumor margins were sufficiently conspicuous and geometrically simple enough to be recognized on all images, and the longest in-plane diameter of the tumor was 10 mm or greater (see Fig 1).

Eleven groups from a diverse set of industry and academic groups participated in the challenge by submitting results from 12 algorithms (one group made two submissions). The participating groups downloaded the images, including the raw image data and location points. The location ("seed") points were defined to lie within the tumor margin. Groups were allowed to select different or multiple seed point(s) for their individual algorithms, provided they used the tumor identification scheme provided. Some of the groups submitted data from the algorithm without any postsegmentation modifications (semiautomated without editing), others submitted data with adjustments made to varying degrees by a reader (semiautomated with editing), and one group submitted both. Each group then uploaded their results using an open-source informatics tool called QI-Bench (37). To establish and maintain anonymity of participants, all communications were handled through the QIBA staff at Radiological Society of North America (RSNA). The participants are as follows (listed alphabetically rather than according to the IDs used in reporting the results of the study): Fraunhofer MEVIS, GE Healthcare, ICON Medical Imaging, KEOSYS, MEDIAN Technologies, Mirada Medical, Perceptive Informatics, Siemens AG, University of California, Los Angeles (UCLA), University of Michigan, and Vital Images.

See the Appendix for detailed algorithm descriptions for each of the participating groups.

### Statistical methods

*Estimation of variability.* The repeatability coefficient (RC) was used to characterize the intra-algorithm variability (6). The RC was defined as

$$\mathrm{RC} = 1.96\sqrt{2\sigma_\varepsilon^2} = 2.77\sigma_\varepsilon,$$

where $\sigma_\varepsilon^2$ is the within-tumor variance. The range in which two measurements on the same tumor were expected to fall for 95% of replicated measurements was given by [−RC, +RC] (38). In this study, we computed the within-tumor variance, and thus RC based on the difference between the test and retest measurements for each algorithm, respectively.

Two calculation methods were used, one using log transformed data and the other a root mean square approach. The root mean square approach proceeds by calculating the square root of the mean of squared tumor-based RC values. Additionally, the within-tumor coefficient of variability ($\mathrm{wCV_{intra}}$) was calculated as a measure of precision for single measurements (6). It was calculated in an analogous fashion by dividing each tumor-based $\sigma_\varepsilon^2$ by the square of the mean of the two measurements and without use of the 2.77 factor. The percent RC (%RC) for an algorithm was determined by multiplying $\mathrm{wCV_{intra}}$ by 2.77. In the logarithmic approach, the %RC is determined by taking an inverse transform. Both $\mathrm{wCV_{intra}}$ and %RC are relative measures proportional to the magnitude of the tumor size. We verified the equivalence of these two methods in a manner described by Bland (39), with the equivalence strongest when the percentage metrics were small. Because we were interested in how the metrics changed for differing tumor sizes, we plotted the percentage metrics as a function of tumor size.

The reproducibility coefficient (RDC), its percentage counterpart percent RDC (%RDC), and $\mathrm{wCV_{inter}}$ were used to characterize inter-algorithm variability (6). The RDC, similar to RC, was calculated from the variance across different algorithm measurements of the same tumor (6). In this study, [−RDC, +RDC] described the range within which approximately 95% of the differences in measurements between two algorithms lie. We reported the reproducibility results in three partitions of algorithms, partitioned based on the intra-algorithm repeatability results. One partition included all algorithms minus the lowest performing algorithm. Another partition included the set of algorithms with %RC less than 30%. A third partition was formed by only including those algorithms with a %RC less than 15%.

A linear mixed effects (LME) model using transformed data was fitted to estimate the relative contributions of different factors to the total variability. The dependent variable in the model was the measured tumor volume. Volume estimation is considered a fixed effect in this model. The independent variables were tumor, algorithm, and tumor-by-algorithm interactions. Model assumptions were evaluated with Q-Q (quantile-quantile) and observed-versus-fitted plots.

*Comparison of segmentation boundaries.* Five groups provided segmentation data in addition to tumor volume measurements, four of which were compatible for analysis (the data from the fifth was submitted with different orientation and scaling). To compare algorithms' segmentation boundaries, we produced a reference segmentation using the simultaneous truth and performance level estimation (STAPLE) method (40) on three-dimensional (3D) volumes. This method performs a voxelwise combination of an arbitrary number of input images, which in our case consisted of the segmentations extracted by the four participant algorithms. Each input segmentation to STAPLE was weighted based on its "performance" as estimated by an expectation-maximization algorithm, described in detail in Rohlfing et al. (41). This algorithm used all input segmentations to create "consensus" results according to the level of overlap among input segmentations. We then compared each individual segmentation result to this reference data. We computed voxelwise accuracy, based on the number of voxels segmented with a particular algorithm compared to the reference data by tabulating counts of true positives (TP, where both the algorithm and the reference contained that voxel), true negative (where neither the algorithm nor the reference contained that voxel), false positive (FP, where the algorithm contained the voxel but the reference

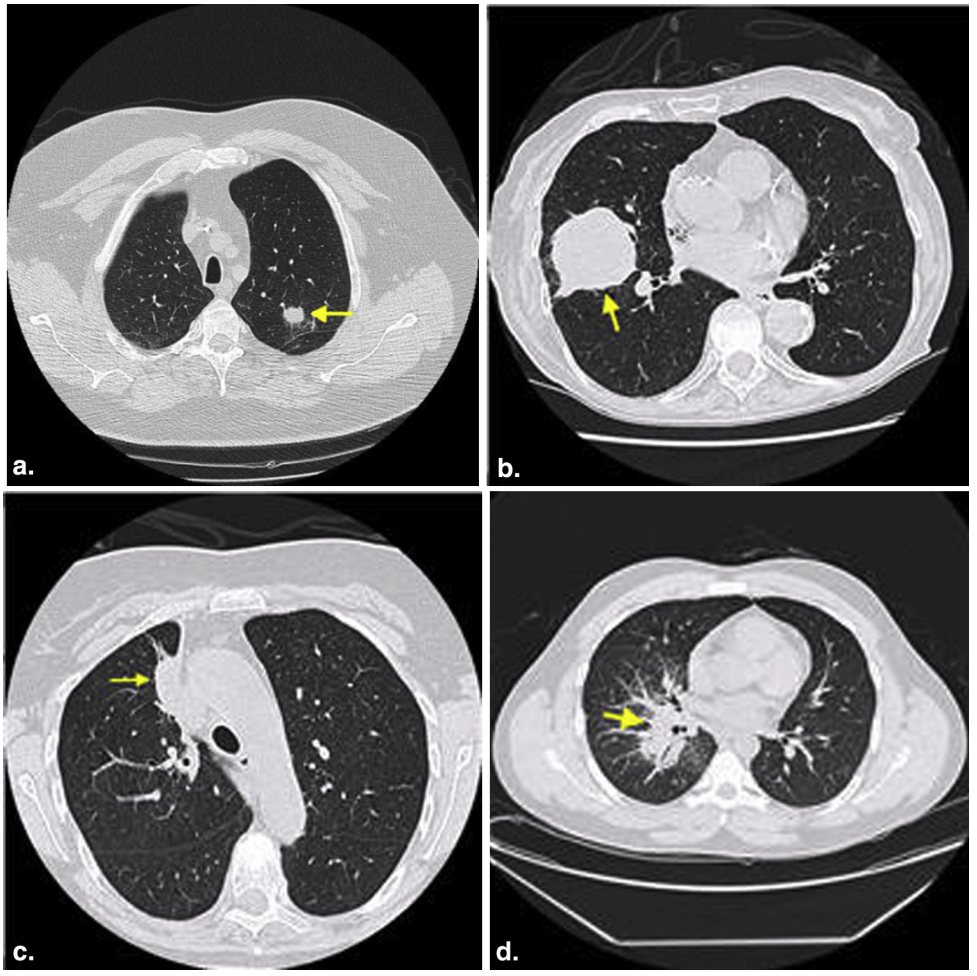


**Figure 1.** Examples of tumors from our study. **(a** and **b)** Examples of tumors that were judged to have met the QIBA measurability criteria, whereas **(c)** and **(d)** were not found to meet the criteria. Image **(c)** was excluded because it demonstrates a large attachment to other pulmonary structures and **(d)** was excluded because it demonstrates a highly invasive structure where the boundary between tumor and nontumor is not well demarcated. QIBA, Quantitative Imaging Biomarker Alliance. (Color version of figure is available online.)

did not), and false negative (FN, where the reference contained the voxel but the algorithm did not). These were used in the calculation of two spatial overlap measures, the Jaccard index (42) and Sørensen-Dice coefficients (43,44) defined as follows:

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \quad \text{Sørensen} - \text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}.$$

The Jaccard index includes a penalty for FP voxels, that is, when the candidate segmentation is larger than the reference segmentation. The Sørensen-Dice coefficient also penalizes FPs, but penalizes more strongly segmentations that have missed TPs. We computed and presented both types of overlap metrics to allow easier and wider comparison to results from other studies.

Excel was used for RC, wCV, and RDC estimation, the R statistical software was used for the mixed effects model, and Matlab was used for overlap metrics.

**TABLE 1. Basic Descriptive Statistics for Measured Tumor Volume**

| Metric | Volume ($mm^3$) | Equivalent Sphere Diameter (mm) |
|---|---|---|
| Arithmetic mean | 24,100 | 36 |
| Geometric mean | 8320 | 25 |
| Median | 9110 | 26 |
| Range | 160,000 | 67 |

## RESULTS

### *Precision of volume measurements*

The total number of possible readings was 744, with each of 12 participating groups submitting both test and retest readings for each of 31 tumors. Of these, 740 were actually submitted, with the following cases missing:

- One group only submitted readings on 30 tumors (rather than 31).

**TABLE 2. Intra-algorithm Repeatability Coefficient (RC) Results**

| | Using all 740 Readings | 34 Anomalous Readings Excluded | | | | |
|---|---|---|---|---|---|---|
| | | All Tumors Pooled | | | Small | Large |
| Group | RC ($mm^3$) | RC ($mm^3$) | %RC | $wCV_{intra}$ | RC ($mm^3$) | RC ($mm^3$) |
| Group 02 | 7557 | 1871 | 13% | 5% | 141 | 1866 |
| Group 03 | 14.060 | 13,568 | 100% | 36% | 1321 | 13,501 |
| Group 04 | 1801 | 1830 | 14% | 5% | 175 | 1825 |
| Group 05 | 3007 | 2177 | 14% | 5% | 245 | 2163 |
| Group 06 | 3418 | 3472 | 20% | 7% | 160 | 3469 |
| Group 07 | 3495 | 3551 | 20% | 7% | 210 | 3545 |
| Group 08 | 2935 | 2982 | 13% | 5% | 147 | 2.979 |
| Group 11 | 41,411 | 39,885 | 50% | 18% | 441 | 39,883 |
| Group 12 | 43,101 | 37,868 | 48% | 18% | 601 | 37,863 |
| Group 14 | 11,081 | 11,259 | 21% | 7% | 161 | 11,257 |
| Group 15 | 2226 | 2261 | 24% | 9% | 321 | 2238 |
| Group 10/16* | 7522 | 7643 | 22% | 8% | 215 | 7639 |

*Volume results submitted under ID Group 16 and segmentation objects submitted under ID Group 10.

- One group only submitted test readings (without retest readings) for two tumors.

Basic descriptive statistics on submitted measurements are given in Table 1, based on the 740 submitted readings. The distribution is skewed because of a very few large reading values, where the mean is much higher than the median.

Detailed review of these 740 submitted readings exposed 34 presumably anomalous readings (leaving 706):

- The unpaired readings were judged anomalous because of having no retest readings.
- Four test-retest reading pairs from three groups differed by log-orders of magnitudes from the rest of the data, suggesting data transcription errors.
- One tumor was particularly challenging for all groups, as judged by the differences in volume measurements being log-orders of magnitudes from each other (whereas other tumors, even other ones that did not otherwise meet the measurability criteria established by QIBA did not exhibit this behavior).

Intra-algorithm repeatability analyses were performed and presented here with and without the readings judged as anomalous. Inter-algorithm reproducibility was assessed with these values excluded. These were removed from the analyses.

***Intra-algorithm repeatability across test-retest repetitions within groups.*** Repeatability results assessed separately for each group are presented in Table 2. Tumors were judged to be "small" if they had a volume of less than 4189 $mm^3$, an equivalent diameter of less than about 20 mm for a sphere, and "large" otherwise (as judged by algorithms individually). Because the algorithm measurements were not normally distributed and did not have constant variance, a log-transformation was applied, reshaping the distribution of the data into a usable form. These summary metrics apply across the large range of tumor volumes included in the study. Figure 2 depicts how the percentage metrics, $wCV_{inter}$, and %RC changed based on the difference between the two measurements for differing tumor sizes, stratified by algorithm performance. Moderately performing algorithms are plotted in the upper panel. In general, these algorithms perform at levels less than 20% RC over most of the range and would be generally understood as being capable of conforming with QIBA repeatability performance requirements. The lower panel depicts the results for the best performing algorithms, which not only provide the best repeatability but could also be considered for interchangeability were they to be used in certain clinical trial designs or clinical use cases.

***Inter-algorithm reproducibility across groups.*** Three separate reproducibility partitions were analyzed. One partition included all groups except Group 3, which demonstrated multiple discrepancies from the behavior exhibited by the other algorithms and had a %RC greater than other groups. Another partition included the set of groups that would be considered to conform to QIBA's requirements as judged by a %RC less than 30%. A third partition was formed by only including those algorithms with a %RC less than 15%. Reproducibility results across all groups are presented in Table 3. Figure 3 depicts how the percentage metrics changed for differing tumor sizes.

***Linear mixed effects model for estimating algorithm versus other sources of error.*** Results of the LMEs are presented in Figure 4, which illustrates the weights of the four different variables on overall volume variability. The variables included in the LME model are tumor, algorithms, and tumor-by-algorithm interactions. Residual error relates to factors not included in the model.

Tumor variation between patients dominates with 96% of total variation, which is expected as this is the component that is attributable to true differences in the object being measured. Tumor-by-algorithm interaction variance

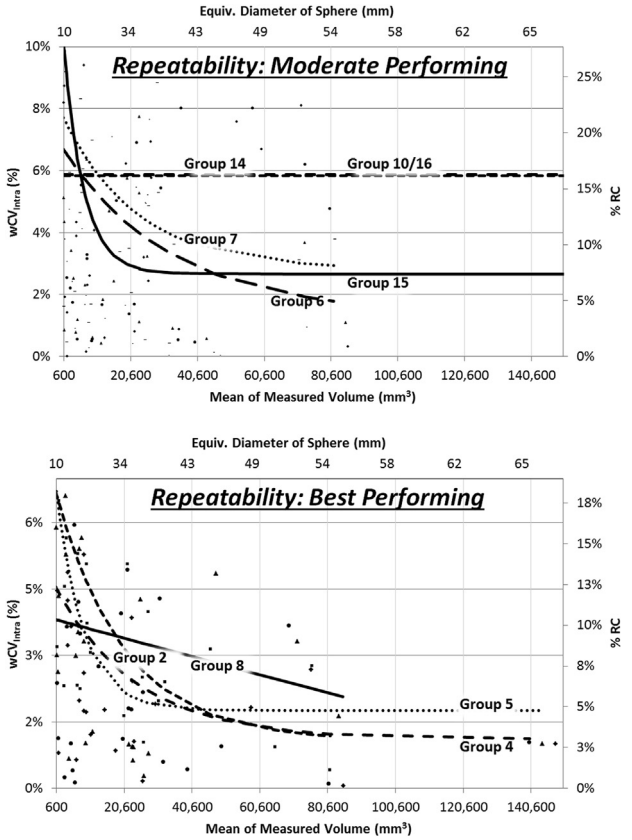


**Figure 2.** Results of intra-algorithm repeatability analysis plotted as a function of measured tumor size. The line fits following exponential functions. Fits for the least performing algorithms could not be made given highly variable results from tumor to tumor. *Upper panel* shows performance with fit lines for moderate performing algorithms, and *lower panel* for best performing algorithms. The fit lines are truncated where they would imply better performance than the sparse set of points at high tumor volumes actually suggest. RC, repeatability coefficient; $wCV_{intra}$, within-tumor coefficient of variability.

comprises the next highest variance, accounting for 3% of the variance, indicating that tumors were measured differently by different algorithms, which is the primary reproducibility result. Residual variance of 1% accounts for factors not attributable to the algorithm performance, for example, hardware variations or scanning technique.

***Stratified reproducibility analyses.*** Four other stratified analyses of reproducibility were carried out, for various combinations of the tumors outlined in Table 4. (For these analyses, definition of small and large was judged based on the average volume estimate for a tumor across the algorithms and using the same 4189 $mm^3$ threshold as used in the repeatability analyses.)

Results for the stratified analyses are summarized in Table 5. The reproducibility of volumetric measurements was better for tumors meeting the QIBA Profile (*Profile = yes*) compared to those tumors that did not (*Profile = no*). This was also reflected in the reduced ratio of algorithm/residual variance for those two analyses. Reproducibility was better when editing was not allowed, indicated by smaller RDC and smaller algorithm/residual variance in the factors model.

**TABLE 3. Inter-algorithm Reproducibility Coefficient (RDC) Results**

| Partition | RDC | %RDC |
|---|---|---|
| All but Group 3 | 25,284 $mm^3$ | 84% |
| Conforming groups | 16,057 $mm^3$ | 70% |
| Best performers | 9290 $mm^3$ | 58% |

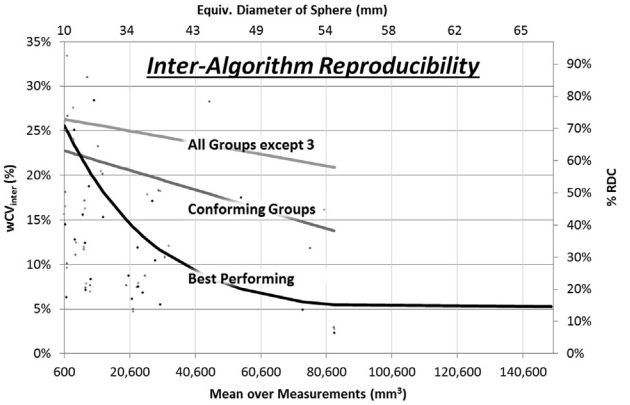


**Figure 3.** Results of inter-algorithm reproducibility analysis plotted across tumor size range. Line fits follow exponential functions. The fit lines are truncated where they would imply better performance than the sparse set of points at high tumor volumes actually suggest. RDC, reproducibility coefficient; $wCV_{intra}$, within-tumor coefficient of variability.

### *Analysis of segmentation boundaries*

Figure 5 shows an example of a reference standard segmentation based on the STAPLE algorithm applied to the segmentation results. A reference segmentation was created for each test-retest repetition and each individual tumor. As indicated in the Materials and Methods section, the reference segmentations were formed using an expectation-maximization algorithm applied to the four compatible submissions. Figure 6 shows an example slice for a single algorithm (Group 08) overlapping with the corresponding reference segmentation. Full evaluation of individual segmentation methods is beyond the scope of the present study, but the detailed maps are provided to the groups who contributed segmentation boundaries for their own analysis.

***Merging and plotting of histograms by metric and group.*** Figure 7 illustrates the histograms of the results created for each group and merged onto a plot that compares the relative segmentation performance of each. The higher number of Sørensen-Dice results greater than 0.8 compared to Jaccard results suggests that oversegmentation (resulting in larger volume measurements) may have been a larger issue than undersegmentation (relative to the imperfect reference standard). Group 10/16 performs best, Group 03 was the least performing algorithm (consistent with its poor computed volume performance), and Groups 04 and 08 depend on the metric used.

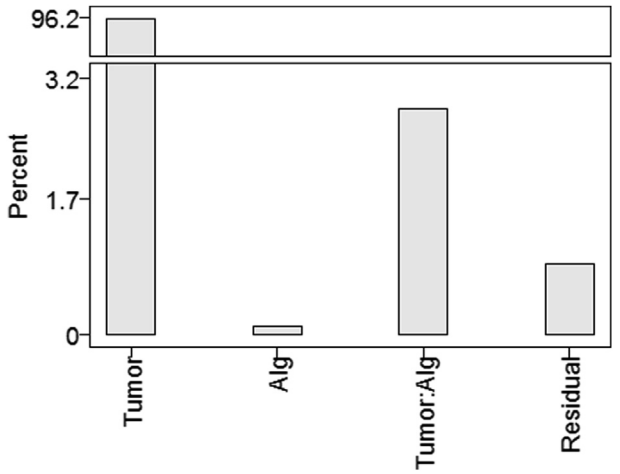


**Figure 4.** Results of LMEs for overall reproducibility analysis, illustrating the percent of total variation captured by each model factor. LMEs, linear mixed effects.

**TABLE 4. Number of Tumors Analyzed in Each Strata**

| Analysis | Strata | N |
|---|---|---|
| Overall | All | 31 |
| | Small | 8 |
| | Large | 23 |
| Profile = yes | All | 20 |
| | Small | 7 |
| | Large | 13 |
| Profile = no | All | 11 |
| | Small | 0 |
| | Large | 11 |
| With editing | All | 31 |
| | Small | 8 |
| | Large | 23 |
| Without editing | All | 31 |
| | Small | 8 |
| | Large | 23 |

Profile = yes or no indicates whether the tumor met the measurability requirements as described previously. With/without editing defines whether postsegmentation contours could be adjusted by a user.

## DISCUSSION

This study was setup to simulate actual practice in the field versus what might be considered from a more controlled academic setting, consistent with QIBA's role of engaging the multiple stakeholders, notably industry, in the practice of quantitative imaging biomarkers such as CT volumetry. In this setting, the information identified in the Appendix is similar to what would be available for methods that are used in practice. Through studies such as ours, we document the performance available, and through the profile writing effort, we seek to identify and reduce sources of variable performance where studies similar to the present one highlight variability. The goal was not to determine the best algorithm but rather the range in performance across diverse algorithms. This is important to the QIBA Profile because the profile describes the performance not of any one algorithm but of a diverse group of algorithms.

**TABLE 5. Summary of Reproducibility Coefficient Results for Stratified Subgroups of Tumors and Algorithms**

| Strata | RDC of Small Tumors | RDC of Large Tumors | Alg/Residual Variance (All Tumors) |
|---|---|---|---|
| Combined | 1290 $mm^3$ | 28,205 $mm^3$ | 3:1 |
| Profile = yes | 1290 $mm^3$ | 6369 $mm^3$ | 2:1 |
| Profile = no | (None in sample) | 41,074 $mm^3$ | 10:2 |
| With editing | 1343 $mm^3$ | 26,760 $mm^3$ | 4:1 |
| Without editing | 1234 $mm^3$ | 33,004 $mm^3$ | 2:1 |

"Alg/Residual Variance" indicates the relative contributions of the two factors to the total variability.

Intra-algorithm %RC ranged from 13% (best performing) to 100% (least performing), with most algorithms demonstrating better percentage performance as the tumor size increased. The four algorithms with the smallest RCs (Groups 2, 4, 5, and 8) were self-identified as semiautomated without editing, whereas the ones with the highest RCs tended to be semiautomated with editing algorithms (Groups 3 and 11, semiautomated with editing) as described in the Appendix. Semiautomated with editing algorithms allow the clinician to correct for egregious segmentation boundaries that can occur when segmenting low-contrast, large, or complex tumors, but this can also introduce the variability often observed from individual perception. One interpretation of these results would be that poorly performing algorithms need editing because of egregious results without it, but once an algorithm is refined to avoid these then editing actually makes the results inferior as they may be best left alone. The algorithms generally show a marked tendency to have smaller percentage metrics (less variability) for larger tumors, which is consistent with the related literature findings (11,45,46). Algorithms were also fairly consistent across tumor sizes, in that the algorithms with the highest wCVs for small tumors also tended to have the highest wCVs for large tumors. The data show some differences; however, for example Group 8 has a lower disparity in wCV*s* between small and large tumors compared to the other best performers.

The RC and wCV results indicate good overall repeatability performance for at least a subset of algorithms, possibly suggesting that some algorithms may also have the potential to be used interchangeably as tumor volume measurement tools for use cases where it is not possible to use a single algorithm. By itself, RC is not sufficient comparing algorithms to unknown truth, motivating the reproducibility analysis, which is a measure of the dispersion in values across algorithms. If the multiple algorithms are individually repeatable but each comes up with (widely) varying measurements, RDC is large (poor) and the algorithms would not be deemed

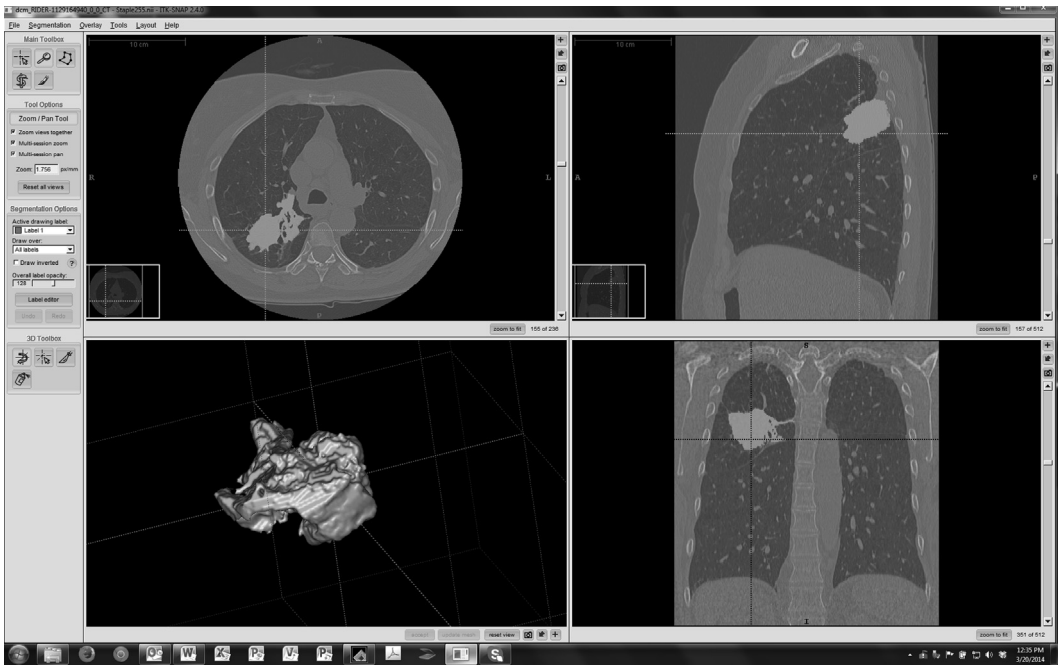

**Figure 5.** Example of a reference truth segmentation (RIDER-1129164940, first repetition, Group 08).

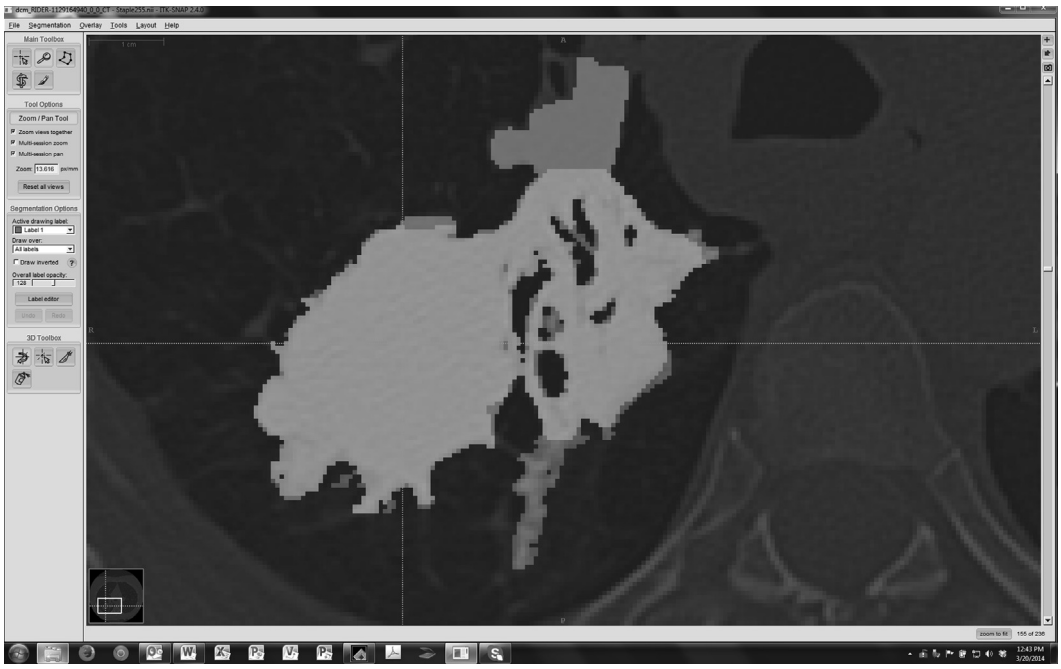

**Figure 6.** Example of a group's result superimposed onto the reference. True positive voxels are rendered as *light gray*, false negative voxels as *dark gray*, and false positive as *medium gray*. True negative pixels are displayed as reduced intensity background image (RIDER-1129164940, first repetition, Group 08).

interchangeable. The only way for RDC to come out small is if the algorithms' measurements are similar among them, and if both the test and retest measurements from each algorithm are included in the calculation of RDC, then it may suffice as a test of interchangeability, hence our approach. Previously reported repeatability results are widely varied across projects and authors; our results demonstrate a range of results as experienced in practice to help account for some of these differences.

The RDC and %RDC were determined in three partitions: 58% for the four best performing groups, 70% for an expanded set of algorithms on the basis of their intra-

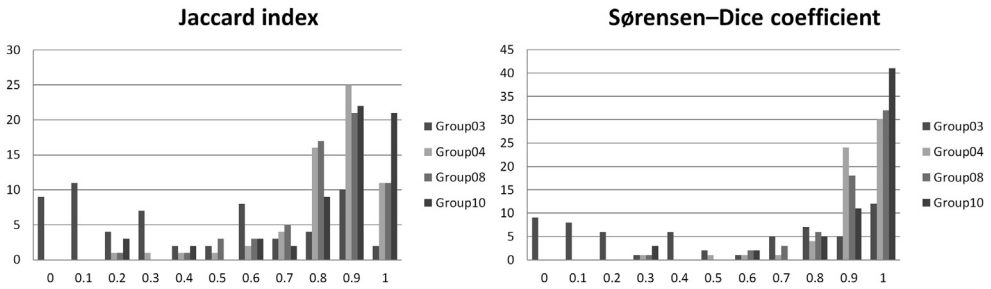


**Figure 7.** Merged histograms for each of three overlap metrics. The x-axis represents the relevant index value (0–1). The y-axis represents the number of tumors with the corresponding index value. Results from four algorithms are plotted with separate colors but combined on each plot to facilitate comparison.

algorithm repeatability being less than 30%, and 84% when all groups except one that was excluded because of erratic behavior. This analysis of the RDC values shows that across all algorithms, the reproducibility performance was low and that, in general, interchanging of all algorithms is not appropriate. This is not surprising because of the low repeatability for some algorithms including Groups 3 and 11 among others. When we evaluated the reproducibility for the subset of algorithms with the best repeatability (eg, Groups 2, 4, 5, and 8), we found that reproducibility improved to 7%. This provides initial evidence that some tumor volume measurement tools might be appropriate for interchangeable use across patient scans acquired at different times. However, this appears to be only possible for a small subset of the algorithms evaluated in this study, and even with these only on tumors with equivalent diameter exceeding 40 mm. For the other algorithms, or for tumors less than 40 mm, care should be taken that the same algorithm is applied at each subsequent time point to eliminate inter-algorithm variability as part of the overall measurement error.

The reproducibility results of Table 5 show that RDC is lowest when algorithms were applied on tumors meeting the measurability criteria defined in the profile as expected. Editing helps performance on larger tumors but no editing is better for small tumors. This may be intuitive, in that larger tumors often include more complex structure, such as larger vessel attachments, and more variation in structure within the tumor whereas smaller tumors might be more easily segmented without need for editing and actually more variable if users try to do so.

Another consideration concerns the extent to which the algorithm may be considered "the end of the line" with respect to variability of the entire process of evaluating tumor size. Our LME analysis showed that more than 96% of the variation is associated with the tumor, leaving just 4% related to other factors. Of this remaining 4%, one-fifth to one-half of this variability comes from sources independent of the algorithms. The ratio of the size of the effect because of algorithm (plus algorithm-tumor interaction) versus the residual informs an "error budget" that may be used for specifying allowable variability because of algorithm versus other parts of the processing chain, so that the system as a whole meets the QIBA claim. On the basis of this, using results summarized in Table 5, not more than two-thirds of the overall variability claim of the system can be allocated to analysis software if the overall system is to meet the QIBA Profile claims. By this measure, conforming algorithms are those with RC less than two-thirds of the overall QIBA Profile claim of 30% or 20%. Eight of the 12 algorithms assessed in this study met this criterion. If the scanner and acquisition parameters are not controlled, demands on algorithms would be much higher. Hence, the QIBA approach is to define performance requirements as means to reduce this variability, although it cannot be eliminated completely.

An additional consideration in characterizing and comparing segmentation algorithms is the segmentation boundaries themselves. We used the Jaccard Index and Sørensen-Dice coefficient for this task. The Jaccard Index and Sørensen-Dice coefficient are consistent across Groups 4 and 8 indicating that the segmentations are generally consistent in both volume and edge profiles for these high RC algorithms. This provides stronger evidence that these two algorithms, and potentially Group 5 as well, could be used interchangeably when evaluating CT tumor progression. Groups 3 and 10/16 did not agree with each other or with Groups 4 and 8 in regard to the Jaccard Index and Sørensen-Dice coefficient indicating that they likely could not be used interchangeably with any other algorithm and may in fact have divergent performance.

The reference standard segmentation was based on the STAPLE algorithm defined across all the four algorithms that provided segmentation results (Groups 3, 4, 8, and 10/16). This is the maximum likelihood segmentation for the tumor based on the segmentations. It may be appealing to think of the reference standard as an estimate for the borders of the true tumor. However, this is generally not appropriate because the segmentation algorithms likely oversegment or undersegment the true tumor, globally or within local regions. Either case would produce a bias in the true boundaries. Even with this limitation, the reference standard can be useful when comparing a set of algorithms because it will show which algorithms have substantial deviation from the norm. This information is likely very helpful in determining which subsets of algorithms can potentially be used interchangeably as discussed previously.

The greatest use of this work and public algorithm challenges in general from a group's point of view or a company seeking to commercialize analysis software for tumor volumetry may be the performance of their algorithm compared to other similar algorithms. Individualized reports inclusive of raw data and intermediate analysis results have been provided to participants in the challenge. The value of the results is highest to those who contributed actual segmentation boundaries, given the ability to distinguish TPs and true negatives from FPs and FNs at a level of granularity allowing algorithm optimization. These data are instrumental to inform the definition of a performance standard for CT tumor volumetry algorithms. Participating groups also benefit, in that algorithm weaknesses are identified.

Our study has limitations. The degree and extent of editing applied to semiautomated algorithms were not held constant between replicates (test-retest measurements), which could have contributed to the overall variability and associated measures of repeatability and reproducibility. Also, our analyses did not account for differences in experience between algorithm operators in terms of interacting with radiological findings or in terms of familiarity/training with the software. Another limitation stems from an explicit determination for this study that workflow should not be constrained, but the related QIBA 1B study suggests that workflow considerations are of substantial importance. In this case, workflow refers to how the repeat scans were processed. In our study, all the scans were processed independently, whereas in part of the QIBA 1B study scans were process in a locked sequential fashion. We had originally thought that semiautomated without editing algorithms (no postsegmentation correction) would not differ in their performance based on workflow, but found that this does not always hold true because ROI and seed placements may be affected. Additionally, the data used in this study were relatively limited, thus only an early version of the QIBA Profile claim specification can be made. Although the data contained an assortment of clinical cases, they did not fully represent the claimed clinical context of use for the corresponding QIBA Profile. Definitive reference data sets that adequately represent the target patient population according to formally assessed statistical criteria should include patients representing a range of common comorbidities, disease characteristics, and imaging settings (eg, sedated vs nonsedated patients). Finally, the manner in which these tests are run and the data collected has implications regarding the interpretation and use of metrics computed and reported. For example, execution of these tests by a trusted third party on sequestered data sets may increase their use.

## ACKNOWLEDGMENTS

The challenge study could not have taken place without the active participation by the organizations that submitted data. We also acknowledge the crucial logistical support from the RSNA staff in administering the application process, which included anonymized interactions among participating study groups, funding provided by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) to defray statistical analysis costs (specifically, this manuscript was prepared in compliance with contract number HHSN268201300071C), and of course the groups themselves as without their effort to produce the submissions there would have been no project.

## REFERENCES

1. Biomarkers Definitions Working Group, . Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001; 69:89–95.
2. Woodcock J, Woosley R. The FDA critical path initiative and its influence on new drug development. Annu Rev Med 2008; 59:1–12.
3. Gurland J, Johnson RO. Case for using only maximum diameter in measuring tumors. Cancer Chemother Rep 1966; 50:119–124.
4. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. Cancer 1976; 38:388–394.
5. Royal HD. Technology assessment: scientific challenges. AJR Am J Roentgenol 1994; 163:503–507.
6. QIBA-Performance-Working-Group. Review of statistical methods for technical performance assessment. Stat Methods Med Res 2015; 24(1): 27–67.
7. Gavrielides MA, Kinnard LM, Myers KJ, et al. Noncalcified lung nodules: volumetric assessment with thoracic CT. Radiology 2009; 251:26–37.
8. Li Q, Gavrielides MA, Zeng R, et al. Volume estimation of low-contrast lesions with CT: a comparison of performances from a phantom study, simulations and theoretical analysis. Phys Med Biol 2015; 60:671.
9. Kinnard LM, Gavrielides MA, Myers KJ, et al. Volume error analysis for lung nodules attached to pulmonary vessels in an anthropomorphic thoracic phantom. Proc SPIE 2008; 6915:69152Q. http://dx.doi.org/10.1117/12.773039.
10. Gavrielides MA, Zeng R, Kinnard LM, et al. A template-based approach for the analysis of lung nodules in a volumetric CT phantom study. Proc SPIE 2009; 7260:726009. http://dx.doi.org/10.1117/12.813560.
11. Winer-Muram HT, Jennings SG, Meyer CA, et al. Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations. Radiology 2003; 229:184–194.
12. Ravenel JG, Leue WM, Nietert PJ, et al. Pulmonary nodule volume: effects of reconstruction parameters on automated measurements—a phantom study. Radiology 2008; 247:400–408.
13. Borradaile K, Ford R. Discordance between BICR readers. Appl Clin Trials, 2010. Nov 1: p. Epub. Available at: http://www.appliedclinicaltrialsonline.com/discordance-between-bicr-readers-0.
14. Gavrielides MA, Zeng R, Myers KJ, et al. Benefit of overlapping reconstruction for improving the quantitative assessment of CT lung nodule volume. Acad Radiol 2013; 20:173–180.
15. Gavrielides MA, Zeng R, Kinnard LM, et al. Information-theoretic approach for analyzing bias and variance in lung nodule size estimation with CT: a phantom study. IEEE Trans Med Imaging 2010; 29:1795–1807.
16. Gavrielides MA, Kinnard LM, Myers KJ, et al. A resource for the assessment of lung nodule size estimation methods: database of thoracic CT scans of an anthropomorphic phantom. Opt Express 2010; 18:15244–15255.
17. Das M, Ley-Zaporozhan J, Gietema HA, et al. Accuracy of automated volumetry of pulmonary nodules across different multislice CT scanners. Eur Radiol 2007; 17:1979–1984.
18. Bolte H, Riedel C, Muller-Hulsbeck S, et al. Precision of computer-aided volumetry of artificial small solid pulmonary nodules in ex vivo porcine lungs. Br J Radiol 2007; 80:414–421.
19. Cagnon CH, Cody DD, McNitt-Gray MF, et al. Description and implementation of a quality control program in an imaging-based clinical trial. Acad Radiol 2006; 13:1431–1441.
20. Goodsitt MM, Chan HP, Way TW, et al. Accuracy of the CT numbers of simulated lung nodules imaged with multi-detector CT scanners. Med Phys 2006; 33:3006–3017.
21. Oda S, Awai K, Murao K, et al. Computer-aided volumetry of pulmonary nodules exhibiting ground-glass opacity at MDCT. AJR Am J Roentgenol 2010; 194:398–406.
22. McNitt-Gray MF, Bidaut LM, Armato SG, et al. Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error. Transl Oncol 2009; 2:216–222.

23. Keil S, Plumhans C, Behrendt FF, et al. Semi-automated quantification of hepatic lesions in a phantom. Invest Radiol 2009; 44:82–88.
24. Gavrielides MA, Li Q, Zeng R, et al. Minimum detectable change in lung nodule volume in a phantom CT study. Acad Radiol 2013; 20:1364–1370.
25. Nietert PJ, Ravenel JG, Leue WM, et al. Imprecision in automated volume measurements of pulmonary nodules and its effect on the level of uncertainty in volume doubling time estimation. Chest 2009; 135: 1580–1587.
26. Prionas ND, Ray S, Boone JM. Volume assessment accuracy in computed tomography: a phantom study. J Appl Clin Med Phys 2010; 11:3037.
27. Chen B, Barnhart H, Richard S, et al. Quantitative CT: technique dependence of volume estimation on pulmonary nodules. Phys Med Biol 2012; 57:1335–1348.
28. Chen B, Barnhart H, Richard S, et al. Volumetric quantification of lung nodules in CT with iterative reconstruction (ASiR and MBIR). Med Phys 2013; 40:111902.
29. Willemink MJ, Leiner T, Budde RP, et al. Systematic error in lung nodule volumetry: effect of iterative reconstruction versus filtered back projection at different CT parameters. AJR Am J Roentgenol 2012; 199:1241–1246.
30. Xie X, Willemink MJ, Zhao Y, et al. Inter- and intrascanner variability of pulmonary nodule volumetry on low-dose 64-row CT: an anthropomorphic phantom study. Br J Radiol 2013; 86:20130160.
31. Linning E, Daqing M. Volumetric measurement pulmonary ground-glass opacity nodules with multi-detector CT: effect of various tube current on measurement accuracy—a chest CT phantom study. Acad Radiol 2009; 16:934–939.
32. Petrick N, Kim HJ, Clunie D, et al. Comparison of 1D, 2D, and 3D nodule sizing methods by radiologists for spherical and complex nodules on thoracic CT phantom images. Acad Radiol 2014; 21:30–40.
33. Buckler AJ, Bresolin L, Dunnick NR, et al. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. Radiology 2011; 258:906–914.
34. CT-Volumetry-Technical-Committee. QIBA profile: CT tumor volume change v2.2 reviewed draft (publicly reviewed version). Available at: http://rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA-CT%20Vol-TumorVolumeChangeProfile_v2.2_ReviewedDraft_08AUG2012.pdf; 2012.
35. McNitt-Gray MF, Kim GH, Zhao B, et al. Determining the variability of lesion size measurements from CT patient data sets acquired under "no change" conditions. Transl Oncol 2015; 8:55–64.
36. Zhao B, James LP, Moskowitz CS, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology 2009; 252:263–272.
37. QI-Bench, free and open-source informatics tooling used to characterize the performance of quantitative medical imaging. Available at: http://www.qi-bench.org/. Accessed June 30, 2013.
38. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets. Transl Oncol 2009; 2:231–235.
39. Bland M. How should I calculate a within-subject coefficient of variation?. cited 2015; Available at: https://www-users.york.ac.uk/~mb55/meas/cv.htm; 2006.
40. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004; 23:903–921.
41. Rohlfing T, Russakoff DB, Maurer CR, Jr. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans Med Imaging 2004; 23: 983–994.
42. Jaccard P. The distribution of the flora in the alpine zone. New Phytol 1912; 11:37–50.
43. Sorensen R. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Nord Med 1948; 40:2389.
44. Dice L. Measures of the amount of ecologic association between species. Ecology 1945; 26:297–302.
45. Reeves AP, Chan AB, Yankelevitz DF, et al. On measuring the change in size of pulmonary nodules. IEEE Trans Med Imaging 2006; 25:435–450.
46. Petrou M, Quint LE, Nan B, et al. Pulmonary nodule volumetric measurement variability as a function of CT slice thickness and nodule morphology. AJR Am J Roentgenol 2007; 188:306–312.

## APPENDIX. ALGORITHM DESCRIPTIONS

Eleven groups participated in the challenge by submitting volume readings for 12 algorithms and five submitted segmentation boundaries, four of which were compatible for analysis. Algorithms from each participating group are described subsequently.

| Participating Group | Description/Workflow |
|---|---|
| Group 02 (volume readings and segmentation boundaries*) | Volumetric analysis was determined using a segmentation approach using a Z-score on the highest conspicuity postcontrast volumetric image set |
| Moderate image/boundary modification (on <50% of the tumors) | A cylinder is placed around the highest conspicuity slice and around all slices above and below this slice in which the tumor is seen |
| | A kernel defined within the region of interest (ROI) is then propagated to other slices using connectivity algorithms. The search is constrained by the predefined cylinder to accelerate the search algorithm |
| Group 03 (volume readings and segmentation boundaries) | One-click user-seeded segmentation |
| | Uses shape and boundary information to delineate the tumor |
| Editing not allowed | The workflow for segmenting lung tumors involves a single click at a seed point roughly centered in the tumor |
| | The algorithm uses the seed point in combination with a thresholded ROI to extract the most probable shape of the tumor |
| Group 04 (volume readings and segmentation boundaries) | Use a trained nonradiologist technician and trained radiologist |
| Limited image/boundary modification (on <15% of the tumors) | As the images would be of chest and the tumors would be in lung parenchyma, all the volume assessments were made using a fixed lung window/level display setting of 200 HU (window) and −1400 HU (level) |
| | Trained nonradiologist opens the images in and uses the tumor location to identify the tumors on images |
| | Trained nonradiologist outlines/ROIs of the identified tumors using automated algorithms |
| | Trained nonradiologist evaluates the quality of the segmentation and adjusts outlines with additional semiautomated tools as necessary |
| | Finally, that image data are submitted to trained radiologist for final assessment of out lines/ROIs. The trained radiologist evaluates the quality of the segmentation and adjusts outlines with automated and semiautomated tools as necessary |
| | Once trained radiologist is satisfied with all the outlines/ROIs of the respective tumors, the automated volume assessment tool is used to calculate volume as volume = (image position interval 1 × area 1) + (image position interval 2 × area 2) ⋯+⋯+ (image position interval n × area n) |
| | The images with ROI is processed, recolored and converted in to .nii file |
| Group 05 (volume readings) | Modelization of the heat-flow between the inside and outside the tumor. On the basis of intensity gradients, in 3D |
| Moderate editing allowed (on <50% of the tumors) | User clicks on a tumor, or draws a diameter joining the boundaries of the tumor => software computes a segmentation of the tumor, and displays its contours |
| | User can then refine the segmentation by the means of a slider => software adjusts the segmentation accordingly, and displays in real-time the new contours |
| | If needed, user can manually edit any contour by drawing it |
| | User finally validates the segmentation => software "locks" the segmentation and extracts the statistics: volume, long axis, short axis, and all intensity-based numbers (average value, standard deviation, and so forth) |
| Group 06 (volume readings) | This algorithm combines the image analysis techniques of region-based active contours and level set approach in a unique way to measure tumor volumes. It may also detect volume changes in part solid and ground glass opacity tumors |
| Editing not allowed (uses only seed points and ROI information) | The user clicks and drags to define an elliptical/circle ROI to initiate the segmentation |
| | The computer then carries out the segmentation, and tumor measurements are saved |

(*Appendix continued*)

| Participating Group | Description/Workflow |
|---|---|
| | The algorithm is an edge-based segmentation method that uniquely combines the image processing techniques of marker-controlled watershed and active contours |
| | An operator initializes the algorithm by manually drawing a region of interest encompassing the tumor on a single slice and then the watershed method generates an initial surface of the tumor in three dimensions, which is refined by the active contours |
| | The volume, maximum diameter, and maximum perpendicular diameter of a segmented tumor are then calculated automatically |
| Group 07 (volume readings) | An initialization sphere is drawn from the center of the mass, on the slice with its largest boundaries, such that it covers the entire extent of the mass. The user determines the center and radius in a single click-drag action, and this initialization circle imposes hard constraints on the maximum boundaries of the 3D segmentation |
| Editing not allowed (uses only seed points and ROI information) | The used algorithm is part of a commercial software package for multimodal oncology treatment assessment and review. Thus, the workflow mimics the typical workflow a user has with this tool: |
| | Select the desired CT data set and load it into any review mode |
| | Select the lung window level setting |
| | Navigate to the tumor center using the pixel and slice locations |
| | Locate the slice where the tumor has the greatest boundaries |
| | Select the algorithm, and initialize the segmentation by clicking in the approximate center of the mass and dragging the mouse to set the radius of the spherical ROI |
| | The spherical ROI contains a fixed inner sphere and the outside sphere, which is set by the mouse dragging motion. The radius is chosen such that the inner circle encompasses most of the mass to be segmented, and the outer sphere can be used as a constraint to prevent any leakage into the chest wall or heart if the mass is attached/abducting to these organs |
| | The computation takes a few seconds (single digit numbers) to compute the result. User may retry the segmentation a few times if the result is unsatisfactory. With each try the previous result is erased, and does not influence the result of preceding try. In this experiment, the user has in overall three tries to get a satisfactory result |
| | Once the segmentation has been determined, the user reads off the volume from the region statistics, which are automatically computed and displayed as soon as the segmentation has been defined. (The volume measurement algorithm counts all voxels whose centroid lies within the segmented contour and multiplies this number with voxel volume) |
| | To document the segmentation result, save the segmentation as an RT-structure set to the data repository |
| Group 08 (volume readings and segmentation boundaries) | Semiautomatic segmentation based on thresholds, growing region, and mathematical morphology processing |
| Moderate editing allowed (on <50% of the tumors) | Digital Imaging and Communications in Medicine (DICOM) images are downloaded and imported into a database. Image data are converted to a proprietary optimized format before the insertion into the database. Tumors' coordinate are downloaded and refor matted by our data manager. Relying on a proprietary validation framework system, landmarks are automatically inserted into the database |
| | The software is allowed then to display the repeated images side by side with the correct landmarks identifying the tumors to segment. The first repetition was edited as a single image. The side-by-side display was available only for the repetition when the first scan edit was locked |
| | Three reviewers are involved, each in charge of segmenting approximately a third of the data set. The data manager made available to the reviewers a commercial semiauto mated algorithm dedicated to lung tumors. Another manual tool can be enabled if semiautomatic segmentations were not fully satisfactory. The data manager recom mended using different window level to better assess tumor boundary, pulmonary |

(*continued on next page*)

(*Appendix continued*)

| Participating Group | Description/Workflow |
|---|---|
| | window level being the major window level to refer to. The data manager recommended correcting semiautomated segmentation as long as the segmentation was not fully satisfactory. Once the whole data set segmented, an additional reviewer was involved to check the whole coherency of the measurements: total number of tumors, no obvious incoherency, correct recording of the data, and so forth |
| | A complete report was extracted. The same validation framework system allowed auto matic extraction of tumors' mask as .mhd format. A third party software as 3D Slicer was used to convert masks to Neuroimaging Informatics Technology Initiative (NIfTI) format |
| Group 11 (volume readings)<br>Editing not allowed (uses only seed points and ROI information) | Method is completely automatic and consists of three steps. First, an ROI is extracted and the tumor is classified as solid or subsolid. In the second step, a binary segmentation mask is computed by an algorithm based on thresholding and morphologic postprocessing, using slightly different procedures for the two classes. Finally, the volume of the tumor is determined by adaptive volume averaging correction |
| | Preprocessing: a stroke is generated from the given center and bounding box by shortening the bounding box diameter to 40% |
| | The segmentation is performed in a cubic ROI, whose edge length is twice the stroke length. The ROI is smoothed with a $3 \times 3$ Gaussian filter and resampled to isotropic voxels and a maximum size of $100 \times 100 \times 100$ voxels. For detecting the tumor type, the local maximum in a $5 \times 5 \times 5$ neighborhood of the ROI center is identified. If its value is greater than $-475$ HU, the tumor is treated as solid, otherwise as subsolid |
| | The ROI center is used as a seed point for region growing. The lower threshold is derived from the 55% quantile of the histogram of the dilated stroke by applying an optimal elliptic function yielding values between $-780$ and $-450$ HU. The resulting mask con tains the complete tumor, but may also leak into adjacent vasculature or, in case of juxtapleural tumors, into structures outside the lungs |
| | To remove vessels, an adaptive opening is used, where the erosion threshold is chosen such that the segmentation has no connection to the ROI boundary anymore. A slight overdilation allows a final refinement of the mask. To avoid leakage outside the lungs, a convex hull of the lung parenchyma is computed within a minimal elliptical region that is fitted to the shape of the tumor. The convex hull is then used as a blocker for the segmentation |
| | Because of the limited spatial resolution of CT and partial volume effects, the volume of a segmented tumor cannot be determined exactly by voxel counting. Instead, voxels in a tube around the segmentation boundary are weighted according to their estimated contribution to the tumor volume. The weight depends on the relation of a voxel's value to the typical tumor and parenchyma densities |
| Group 12 (volume readings)<br>Moderate editing allowed (on <50% of the tumors) | We start with an automatic method (submitted Group 11) and correct results interactively if necessary. The user draws partial contours, which are included in the segmentation in the edited slice. Additionally, the correction is automatically propagated to a set of neighboring slices by sampling the contour, matching points to the next slice, and connecting them with a live-wire method |
| | Interactive correction: our interactive correction tool provides an efficient way to fix seg mentation results, which are mostly correct but need some refinement. The user draws partial contours indicating the desired segmentations, which are then automatically p ropagated into 3D. Seed points calculated from the user contour are moved to adjacent slices by a block matching algorithm and the seed points are connected by a live-wire algorithm. For the submission, correction was performed by two experienced developers in consensus |
| | Volumetry: the volumetry used for automatic results is integrated in the segmentation algorithm. To ensure consistency after interactive correction, the change in the number of voxels is computed and multiplied with the (partial-volume-corrected) volume of the initial result |

(*Appendix continued*)

| Participating Group | Description/Workflow |
|---|---|
| Group 14 (volume readings)<br>Editing not allowed (uses only seed points and ROI information) | The system is fully automated after manual input of an approximate bounding box for the tumor of interest. Within the bounding box, the system automatically processes the images in three stages—preprocessing, initial segmentation, and 3D level set segmentation<br>In the first stage, a set of smoothed images and a set of gradient images are obtained by using 3D preprocessing techniques to the original CT images. Smoothing, anisotropic diffusion, gradient filtering, and rank transform of the gradient magnitude are used to obtain a set of edge images<br>In the second stage, based on attenuation, gradient, and location, a subset of pixels is selected, which are relatively close to the center of the tumor and belong to smooth (low gradient) areas. The pixels are selected within an ellipsoid that has axis lengths one-half of those of the inscribed ellipsoid within the bounding box. This subset of pixels is considered to be a statistical sample of the full population of pixels in the tumor. The mean and standard deviation of the intensity values of the pixels belonging to the subset are calculated. The preliminary tumor contour is obtained after thresholding and includes the set of pixels falling within three SDs of the mean and with values greater than the fixed background threshold. A morphologic dilation filter, a 3D flood fill algorithm, and a morphologic erosion filter are used to the contour to connect the nearby components and extract an initial segmentation surface. The size of the ellipsoid and the remaining parameters are selected experimentally in a way that enables segmentation of a variety of tumors, including necrotic tumors<br>In the third stage, the initial segmentation surface is propagated by using a 3D level set method. Four level sets are applied sequentially to the initial contour. The first three level sets are applied in 3D with a predefined schedule of parameters, and the last level set is applied in 2D to every section of the resulting 3D segmentation to obtain the final con tour. The first level set slightly expands and smoothes the initial contour. The second level set pulls the contour toward the sharp edges, but at the same time, it expands slightly in regions of low gradient. The third level set further draws the contour toward the sharp edges. The 2D level set performs final refinement of the segmented contour on every section |
| Group 15 (volume readings)<br>Moderate editing allowed (on <50% of the tumors) | The software used is essentially a semiautomated contouring method. The user clicks on a voxel located inside the tumor of interest and then drags a line to the outside the tumor (to the background)<br>The voxels along that line are sampled and a histogram of intensities (Hounsfield Units) is created<br>A statistical method is used to determine the threshold that best separates the two distributions (tumor and background) in that histogram<br>Once that threshold is determined, the software uses a 3D (or if selected a 2D) seeded region growing using the initial voxel selected as the point inside the tumor and the threshold determined from the histogram analysis<br>The tool also provides several user editing tools such as adding and erasing voxels from the contour, and so forth. The workflow description is as follows:<br>Each contour is automatically stored in a database linked to the experiment along with meta data such as patient ID, contouring individual's ID, and so forth. Each contoured object has a unique ID that is linked to the series UID to maintain its identity<br>Once the contour is completed and accepted, the volume of the contoured object is calculated. This is done essentially by counting the number of voxels within the boundaries of the contoured object and multiplying that by the voxel size (as derived from DICOM header data) |
| Group 10/16 (volume readings and segmentation boundaries[†]) | As the input for the algorithm, the user has to draw a stroke being favorably the largest diameter in the axial orientation or click a point in the given lung tumor. Usually, the decision to use a stroke or a single click point depends on the size of the tumor to be |

(*continued on next page*)

(*Appendix continued*)

| Participating Group | Description/Workflow |
|---|---|
| Limited editing allowed (on <50% of the tumors) | segmented (for bigger tumors, a stroke is preferable, whereas for small tumors, a single click is sufficient)<br>In the next step, a volume of interest (VOI) around the tumor is estimated. In the case where the algorithm has been initialized with stroke, the size of the VOI depends on the length of the stroke<br>3D region growing is conducted in a VOI starting from seeds generated along the stroke or around the click point, depending on the initialization<br>Adjacent structures of similar density (pleura, vessels) are separated by a set of inter changing morphologic operations (erosion, dilation, convex hull, and binary combination with region growing mask.)<br>Finally, a plausibility check between the resulting segmentation mask and the position of the initial stroke or click point is conducted. If necessary, initial thresholds are read justed and the whole procedure (steps 2–5) is repeated<br>For the case when the semiautomatic results are not satisfactory, the software provides the possibility of correcting the results by drawing contours in selected slices and then propagating the contours in an automatic manner onto the whole 3D segmentation. The algorithm performs best optimally for the resolution up to 2 mm, although it still works reasonably well for thicker slices such as 5 mm |

Three groups (Groups 01, 09, and 13) initially applied but did not submit results.

*Alignment issues prevented inclusion in the segmentation boundary analysis.

†Volume results submitted under ID Group 16 and segmentation objects submitted under ID Group 10.