

QIBA Profile:

CT Tumor Volume Change (CTV-1)

Version 2.4-ish

2016-Feb-13
Status: Reviewed Draft (Public Comments Addressed)

Note to users – when referencing this QIBA Profile document, please use the following format:

CT Volumetry Technical Committee. CT Tumor Volume Change Profile, Quantitative Imaging Biomarkers Alliance. Version 2.4. Reviewed Draft. QIBA, May 1, 2015. Available at: http://rsna.org/QIBA\_.aspx

Table of Contents

[1. Executive Summary 4](#_Toc443126679)

[2. Clinical Context and Claim(s) 5](#_Toc443126680)

[3. Profile Requirements 7](#_Toc443126681)

[3.1. Subject Handling 9](#_Toc443126682)

[3.1.1 Discussion 9](#_Toc443126683)

[3.1.2 Specification 11](#_Toc443126684)

[3.2. Image Data Acquisition 12](#_Toc443126685)

[3.2.1 Discussion 12](#_Toc443126686)

[3.2.2 Specification 14](#_Toc443126687)

[3.3. Image Data Reconstruction 15](#_Toc443126688)

[3.3.1 Discussion 15](#_Toc443126689)

[3.3.2 Specification 17](#_Toc443126690)

[3.4. Image QA 18](#_Toc443126691)

[3.4.1 Discussion 18](#_Toc443126692)

[3.4.2 Specification 20](#_Toc443126693)

[3.5. Image Analysis 20](#_Toc443126694)

[3.5.1 Discussion 20](#_Toc443126695)

[3.5.2 Specification 23](#_Toc443126696)

[4. Assessment Procedures 25](#_Toc443126697)

[4.1. Assessment Procedure: In-plane Spatial Resolution 25](#_Toc443126698)

[4.2. Assessment Procedure: Voxel Noise 26](#_Toc443126699)

[4.3. Assessment Procedure: Tumor Volume Computation 27](#_Toc443126700)

[4.4. Assessment Procedure: Tumor Volume Change Repeatability 27](#_Toc443126701)

[4.4.1 obtain test image set 28](#_Toc443126702)

[4.4.2 determine volume change 29](#_Toc443126703)

[4.4.3 calculate statistical metrics of performance 29](#_Toc443126704)

[4.5. Assessment Procedure: Tumor Volume Bias and Linearity 29](#_Toc443126705)

[4.5.1 obtain test image set 30](#_Toc443126706)

[4.5.2 determine volume change 31](#_Toc443126707)

[4.5.3 calculate statistical metrics of performance 31](#_Toc443126708)

[4.X. Assessment Procedure: Image Acquisition Site Performance 32](#_Toc443126709)

[4.X.1 Acquisition Validation 33](#_Toc443126710)

[4.X.2 Test Image Set 33](#_Toc443126711)

[Closed Issues: 34](#_Toc443126712)

[Appendices 38](#_Toc443126713)

[Appendix A: Acknowledgements and Attributions 38](#_Toc443126714)

[Appendix B: Background Information 39](#_Toc443126715)

[B.1 QIBA 39](#_Toc443126716)

[B.2 CT Volumetry for Cancer Response Assessment: Overview and Summary 40](#_Toc443126717)

[B.3 Detailed Literature Review by Indication 44](#_Toc443126718)

[Lung Cancer (Tables B.2 and B.3) 44](#_Toc443126719)

[Primary Liver Cancer and Metastatic Lesions in the Liver (Table B.4) 46](#_Toc443126720)

[Lymphoma (Table B.5) 47](#_Toc443126721)

[Colorectal and Gastric Cancers (Table B.6) 48](#_Toc443126722)

[Head and Neck Cancer (Table B.7) 48](#_Toc443126723)

[Sarcoma (Table B.8) 49](#_Toc443126724)

[B.4 References 54](#_Toc443126725)

[Appendix C: Conventions and Definitions 60](#_Toc443126726)

[Appendix D: Model-specific Instructions and Parameters 61](#_Toc443126727)

[Appendix E: Metrology Definitions and Methods 64](#_Toc443126728)

[E.2 Considerations for Performance Assessment of Tumor Volume Change 66](#_Toc443126729)

# 1. Executive Summary

The goal of a QIBA Profile is to help achieve a useful level of performance for a given biomarker.

The biomarker performance is described in the Claim (See Section 2) and the profile specifies the necessary behaviors for a set of actors participating in activities that contribute to generating the biomarker (See Section 3). Some requirements are evaluated using specific Assessment Procedures (See Section 4).

This QIBA Profile (CT Tumor Volume Change) addresses tumor volume change which is often used as a biomarker of disease progression or response to treatment. It places requirements on actors (Acquisition Devices, Technologists, Radiologists, Reconstruction Software and Image Analysis Tools) involved in activities (Subject Handling, Image Data Acquisition, Image Data Reconstruction, Image QA and Image Analysis). The requirements are primarily focused on achieving sufficient accuracy and avoiding unnecessary variability of the tumor volume measurements.

The clinical performance target is to achieve a 95% confidence interval for the tumor volume change with precision of -25% to +30%.

This document is intended to help clinicians basing decisions based on these measurements, imaging staff generating these measurements, manufacturer staff developing related products, purchasers of such products and investigators designing trials with imaging endpoints.

Note that this document are only states requirements to achieve the claim, not “requirements on standard of care.” Specifically, meeting the goals of this Profile is secondary to properly caring for the patient.

QIBA Profiles addressing other imaging biomarkers using CT, MRI, PET and Ultrasound can be found at qibawiki.rsna.org.

# 2. Clinical Context and Claim(s)

****Clinical Context****

Quantifying the volumes of tumors and measuring tumor longitudinal changes within subjects (i.e. evaluating growth or regression with image processing of CT scans acquired at different timepoints).

**Compliance with this Profile by all relevant staff and equipment supports the following claim(s):**

Claim 1: A measured increase in mass volume of 30% or more indicates that a true increase has occurred with 95% confidence.

Claim 2:  For a measured x% change in tumor volume of x%, a 95% confidence interval for the true change in percent is [x-25, x+30].

**This claim holds when:**

* **the tumor is measurable at both timepoints (i.e., tumor margins are sufficiently conspicuous and geometrically simple enough to be recognized on all images in both scans; the tumor is unattached to other structures of equal density)**
* **the tumor longest in-plane diameter is between 10 mm (volume 0.5 cm3) and 100 mm (volume 524 cm3) at both timepoints**

Discussion

The -25% and +30% boundaries can be thought of as “error bars” or “noise” around the measurement of volume change. If you measure change within this range, you cannot be certain that there has really been a change. However, if a tumor changes size beyond these limits, you can be 95% confident there has been a true change in the size of the tumor, and the perceived change is not just measurement variability. Note that this does not address the biological significance of the change, just the likelihood that the measured change is real.

Clinical interpretation (progression/response):
To be 95% confident there has been a true increase in tumor volume, the measured change should be at least yy%. To be 95% confident there has been a true decrease in tumor volume, the measured change should be at least -xx%. This is based on the 95% confidence interval of [-xx%, yy%] in the Claim. Whether such a true increase or decrease in tumor volume constitutes clinically meaningful disease progression or response is a distinct decision that requires a clinician’s judgment. Note that there are currently no validated response criteria based on volume (RECIST is for unidimensional measurements).

Clinical interpretation (magnitude of change): The magnitude of the true change is defined by the measured change and the error bars (+-83%). If you measure the volume to be 200mm3 at baseline and 380mm3 at follow-up, then the measured change is a 90% increase in volume (i.e., 100x(380-200)/200). The 95% confidence interval for the true change is a 7% to 173% increase in volume.

The asymmetric range in Claim 1 (-25% to +30%) is due to the way change is conventionally expressed and how measurements are performed.

The lower bound on the tumor longest in-plane diameter is set to limit the variability introduced when approaching the resolution of the dataset, e.g. partial volume. The upper bound is set to limit the variability introduced by more complex tumor morphology and organ involvement, and also to keep performance assessment procedures manageable.

While Claim 1 has been informed by an extensive review of the literature and expert consensus that has not yet been fully substantiated by studies that strictly conform to the specifications given here. The expectation is that during field test, data on the actual field performance will be collected and any appropriate changes made to the claim or the details of the Profile. At that point, this caveat may be removed or re-stated.

The performance values in Claim 1 reflect the likely impact of variations permitted by this Profile. The Profile permits different compliant actors (acquisition device, radiologist, image analysis tool, etc.) at the two timepoints (i.e. it is not required that the same scanner or image analysis tool be used for both exams of a patient). If one or more of the actors are the same, the implementation is still compliant with this Profile and it is expected that the measurement performance will be improved. To give a sense of the possible improvement, the following table presents expected precision for alternate scenarios, however except for the leftmost, these precision values are **not** Claims of this Profile.

Table 1: Expected Precision for Alternate Scenarios (Informative)

|  |  |
| --- | --- |
| 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 |
| **47%** | 46% | 33% | 32% | 38% | 36% | 13% | 11% |

Notes:

1. Precision is expressed here as the total deviation index.

2. 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.

3. A 95% confidence interval for the magnitude of the true change is given by: ± the relevant precision value

# 3. Profile Requirements

The Profile is documented in terms of “Actors” performing “Activities”.

Equipment, software, staff or sites may claim conformance to this Profile as one or more of the “Actors” in the following table. Compliant Actors shall support the listed Activities by meeting all requirements in the referenced Section. Failing to comply with a “shall” is a protocol deviation. Although deviations invalidate the Profile Claim, such deviations may be reasonable and unavoidable as discussed below.

Table 1: Actors and Required Activities

|  |  |  |
| --- | --- | --- |
| **Actor** | **Activity** | **Section** |
| Acquisition Device | Subject Handling | 3.1. |
| Image Data Acquisition | 3.2. |
| Technologist | Subject Handling | 3.1. |
| Image Data Acquisition | 3.2. |
| Image Data Reconstruction | 3.3. |
| Radiologist | Subject Handling | 3.1. |
| Image QA | 3.4. |
| Image Analysis | 3.5. |
| Reconstruction Software | Image Data Reconstruction | 3.3. |
| Image Analysis Tool | Image Analysis | 3.5. |

For the Acquisition Device, Reconstruction Software and Image Analysis Tool actors, while it will typically be the manufacturer who claims the actor is conformant, it is certainly possible for a site to run the necessary tests/checks to confirm compliance and make a corresponding claim. This might happen in the case of an older model device which the manufacturer is no longer promoting, but which a site needs a compliance claim to participate in a clinical trial.

The sequencing of the Activities specified in this Profile are shown in Figure 1:

Acq.

*Subtract*

*volumes*

Subject

Handling

Recon

*Obtain images per timepoint (2)*

*Imaging*

*Agent*

(

*if any*

)

images

*Measure change per target lesion*

*Measure change in target lesion volume*

*Calculate*

*volume*

*Calculate*

*volume*

volume

changes

*volumes*

*...*

 QA

 Image Analysis

Figure 1: CT Tumor Volumetry - Activity Sequence

The method for measuring change in tumor volume may be described as a pipeline. Subjects are prepared for scanning, raw image data is acquired, images are reconstructed and evaluated. Such images are obtained at two (or more) time points. Image analysis assesses the degree of change between two time points for each evaluable target tumor by calculating absolute volume at each time point and subtracting. Volume change is expressed as a percentage (delta volume between the two time points divided by the average of the volume at time point 1 and time point t).

The change may be interpreted according to a variety of different response criteria. These response criteria are beyond the scope of this document. Detection and classification of tumors as target is also beyond the scope of this document.

The Profile does not intend to discourage innovation, although it strives to ensure that methods permitted by the profile requirements will result in performance that meets the Profile Claim. The above pipeline provides a reference model. Algorithms which achieve the same result as the reference model but use different methods may be permitted, for example by directly measuring the change between two image sets rather than measuring the absolute volumes separately. Developers of such algorithms are encouraged to work with the appropriate QIBA committee to conduct any groundwork and assessment procedure revisions needed to demonstrate the requisite performance.

The requirements included herein are intended to establish a baseline level of capabilities. Providing higher performance or advanced capabilities is both allowed and encouraged. The Profile does not intend to limit how equipment suppliers meet these requirements.

This Profile is “lesion-oriented”. The Profile requires that images of a given tumor be acquired and processed the same way each time. It does not require that images of tumor A be acquired and processed the same way as images of tumor B; for example, tumors in different anatomic regions may be imaged or processed differently, or some tumors might be examined at one contrast phase and other tumors at another phase.

The requirements in this Profile do not codify a Standard of Care; they only provide guidance intended to achieve the stated Claim. Although deviating from the specifications in this Profile may invalidate the Profile Claims, the radiologist or supervising physician is expected to do so when required by the best interest of the patient or research subject. How study sponsors and others decide to handle deviations for their own purposes is entirely up to them.

Since much of this Profile emphasizes performing subsequent scans consistent with the baseline scan of the subject, the parameter values chosen for the baseline scan are particularly significant and should be considered carefully.

In some scenarios, the “baseline” might be defined as a reference point that is not necessarily the first scan of the patient.

## 3.1. Subject Handling

### 3.1.1 Discussion

This Profile will refer primarily to “subjects”, keeping in mind that the requirements and recommendations apply to patients in general, and subjects are often patients too.

**Timing Relative to Index Intervention Activity**

When the Profile is being used in the context of a clinical trial, refer to relevant clinical trial protocol for further guidance or requirements on timing relative to index intervention activity.

**Timing Relative to Confounding Activities**

This document does not presume any timing relative to other activities.

Fasting prior to a contemporaneous FDG PET scan or the administration of oral contrast for abdominal CT is not expected to have any adverse impact on this Profile.

**Contrast Preparation and Administration**

Contrast characteristics influence the appearance, conspicuity, and quantification of tumor volumes.
Non-contrast CT might not permit an accurate characterization of the malignant visceral/nodal/soft-tissue tumors and assessment of tumor boundaries.

However, the **use of contrast** material (intravenous or oral) may not be medically indicated in defined clinical settings or may be contra-indicated for some subjects. It is up to Radiologists and supervising physicians to determine if the contrast protocol is appropriate for the subject. They may omit intravenous contrast or vary administration parameters when required by the best interest of patients or research subjects, in which case tumors may still be measured but the measurements will not be subject to the Profile claims.

It is important that the **Contrast Protocol** achieves a consistent phase and degree of enhancement. Bolus tracking is a good tool if available, but is not required. When using bolus tracking, be consistent between scans with where the ROI used for triggering is placed and the threshold used to trigger the scan. When bolus tracking is not available, be consistent between the scans with the contrast volume, rate, scan timing after injection, and use (or lack) of a saline flush. The use of oral contrast material should be consistent for all abdominal imaging timepoints.

Recording the use and type of contrast, actual dose administered, injection rate, and delay in the image header by the Acquisition Device is recommended. This may be by automatic interface with contrast administration devices in combination with text entry fields filled in by the Technologist. Alternatively, the technologist may enter this information manually on a form that is scanned and included with the image data as a DICOM Secondary Capture image.

**Subject Positioning**

Positioning the subject Supine/Arms Up/Feet First has the advantage of promoting consistency (if it’s always the same, then it’s always consistent with baseline), and reducing cases where intravenous lines go through the gantry, which could introduce artifacts. Consistent positioning avoids unnecessary changes in attenuation, changes in gravity induced shape and fluid distribution, or changes in anatomical shape due to posture, contortion, etc. Significant details of subject positioning include the position of their arms, the anterior-to-posterior curvature of their spines as determined by pillows under their backs or knees, the lateral straightness of their spines. Prone positioning is not recommended.

When the patient is supine, the use of positioning wedges under the knees and head is recommended so that the lumbar lordosis is straightened and the scapulae are both in contact with the table. However, the exact size, shape, etc. of the pillows is not expected to significantly impact the Profile Claim. It is expected that clinical trial documentation or local clinical practice will specify their preferred patient positioning.

Recording the Subject Positioning and Table Heights in the image header is helpful for auditing and repeating baseline characteristics.

Artifact sources, in particular metal and other high density materials, can degrade the reconstructed volume data such that it is difficult to determine the true boundary of a tumor. Due to the various scan geometries, artifacts can be induced some distance from the artifact source. The simplest way to ensure no degradation of the volume data is to remove the artifact sources completely from the patient during the scan, if feasible. Although artifacts from residual oral contrast in the esophagus could affect the measurement of small tumors near the esophagus, this is not addressed here.

Consistent centering of the patient avoids unnecessary variation in the behavior of dose modulation algorithms during scan.

**Instructions to Subject During Acquisition**

Breath holding reduces motion that might degrade the image. Full inspiration inflates the lungs, which separates structures and makes tumors more conspicuous.

Since some motion may occur due to diaphragmatic relaxation in the first few seconds following full inspiration, a proper breath hold will include instructions like "Lie still, breathe in fully, hold your breath, and relax”, allowing 5 seconds after achieving full inspiration before initiating the acquisition.

Although performing the acquisition in several segments (each of which has an appropriate breath hold state) is possible, performing the acquisition in a single breath hold is likely to be more easily repeatable and does not depend on the Technologist knowing where the tumors are located.

**Timing/Triggers**

The amount and distribution of contrast at the time of acquisition can affect the appearance and conspicuity of tumors.

### 3.1.2 Specification

| **Parameter** | **Actor** | **Specification** |
| --- | --- | --- |
| Use of intravenous contrast  | Radiologist | Shall determine whether the selected contrast protocol, if any, will achieve sufficient tumor conspicuity  |
| Technologist | Shall use intravenous contrast parameters consistent with baseline.Shall document the total volume of contrast administered, the concentration, the injection rate, and whether a saline flush was used.  |
| Contrast Protocol | Technologist | Shall use a contrast protocol that achieves enhancement consistent with baseline |
| Use of oral contrast | Technologist | Shall use oral contrast parameters consistent with baseline.Shall document the total volume of contrast administered and the type of contrast.  |
| Subject Positioning | Technologist | Shall position the subject consistent with baseline. If baseline positioning is unknown, position the subject Supine if possible, with devices such as positioning wedges placed as described above. |
| Artifact Sources | Technologist | Shall remove or position potential sources of artifacts (specifically including breast shields, metal-containing clothing, EKG leads and other metal equipment) such that they will not degrade the reconstructed CT volumes. |
| Table Height & Centering | Technologist | Shall adjust the table height for the mid-axillary plane to pass through the isocenter. Shall position the patient such that the “sagittal laser line” lies along the sternum (e.g. from the suprasternal notch to the xiphoid process). |
| Breath hold | Technologist | Shall instruct the subject in proper breath-hold and start image acquisition shortly after full inspiration, taking into account the lag time between full inspiration and diaphragmatic relaxation. Shall ensure that for each tumor the breath hold state is consistent with baseline. |
| Image Header | Technologist | Shall record factors that adversely influence subject positioning or limit their ability to cooperate (e.g., breath hold, remaining motionless, agitation in subjects with decreased levels of consciousness, subjects with chronic pain syndromes, etc.).  |
| Acquisition Device | Shall provide corresponding data entry fields. |
| Contrast-based Acquisition Timing | Technologist | Shall ensure that the time-interval between the administration of intravenous contrast (or the detection of bolus arrival) and the start of the image acquisition is consistent with baseline (i.e. obtained in the same phase; arterial, venous, or delayed).Shall ensure that the time-interval between the administration of oral contrast and the start of the image acquisition is consistent with baseline. (Note that the tolerances for oral timing are larger than for intravenous). |
| Image Header | Acquisition Device | Shall record actual timing and triggers in the image header by including the Contrast/Bolus Agent Sequence (0018,0012). |

## 3.2. Image Data Acquisition

### 3.2.1 Discussion

CT scans for tumor volumetric analysis can be performed on any equipment that complies with the specifications set out in this Profile. However, we strongly encourage performing all CT scans for an individual subject on the same platform (manufacturer, model and version) which we expect will further reduce variation.

Many scan parameters can have direct or indirect effects on identifying, segmenting and measuring tumors. To reduce this potential source of variance, all efforts should be made to have as many of the scan parameters as possible consistent with the baseline.

**Consistency with the baseline** implies a need for a method to record and communicate the baseline settings and make that information available at the time and place that subsequent scans are performed. Although it is conceivable that the scanner could retrieve prior/baseline images and extract acquisition parameters to encourage consistency, such interoperability mechanisms are not defined or mandated here and cannot be depended on to be present or used. Similarly, managing and forwarding the data files when multiple sites are involved may exceed the practical capabilities of the participating sites. Sites should be prepared to use manual methods instead.

The goal of parameter consistency is to achieve consistent performance. Parameter consistency when using the same scanner make/model generally means using the same values. Parameter consistency when the baseline was acquired on a different make/model may require some “interpretation” to achieve consistent performance since the same values may produce different behavior on different models. The parameter sets in Appendix D may be helpful in this task.

The approach of the specifications here, and in the reconstruction section, is to focus as much as possible on the characteristics of the resulting dataset, rather than one particular technique for achieving those characteristics. This is intended to allow as much flexibility as possible for product innovation and reasonable adjustments for patient size (such as increasing acquisition mAs and reconstruction DFOV for larger patients), while reaching the performance targets. Again, the technique parameter sets in Appendix D may be helpful for those looking for more guidance.

The purpose of the minimum **scan duration** requirement is to permit acquisition of an anatomic region in a single breath-hold, thereby preventing respiratory motion artifacts or anatomic gaps between breath-holds. This requirement is applicable to scanning of the chest and upper abdomen, the regions subject to these artifacts, and is not required for imaging of the head, neck, pelvis, spine, or extremities.

**Coverage** of additional required anatomic regions (e.g. to monitor for metastases in areas of likely disease) depends on the requirements of the clinical trial or local clinical practice. In baseline scans, the tumor locations are unknown and may result in a tumor not being fully within a single breath-hold, making it “unmeasurable” in the sense of this Profile.

**Pitch** is chosen so as to allow completion of the scan in a single breath hold.

For subjects needing two or more breath-holds to fully cover an anatomic region, different tumors may be acquired on different breath-holds. It is still necessary that each tumor be fully included in images acquired within a single breath-hold to avoid discontinuities or gaps that would affect the measurement.

**Scan Plane** (transaxial is preferred) may differ between subjects due to the need to position for physical deformities or external hardware. For an individual subject, a consistent scan plane will reduce unnecessary differences in the appearance of the tumor.

**Total Collimation Width** (defined as the total nominal beam width, NxT, for example 64x1.25mm) is often not directly visible in the scanner interface. Manufacturer reference materials typically explain how to determine this for a particular scanner make, model and operating mode. Wider collimation widths can increase coverage and shorten acquisition, but can introduce cone beam artifacts which may degrade image quality. Imaging protocols will seek to strike a balance to preserve image quality while providing sufficient coverage to keep acquisition times short.

**Nominal Tomographic Section Thickness** (T), the term preferred by the IEC, is sometimes also called the Single Collimation Width. It affects the spatial resolution along the subject z-axis.

Smaller voxels are preferable to reduce partial volume effects and provide higher accuracy due to higher spatial resolution. The resolution/voxel size that reaches the analysis software is affected by both acquisition parameters and reconstruction parameters.

X-ray CT uses ionizing radiation. Exposure to radiation can pose risks; however as the radiation dose is reduced, image quality can be degraded. It is expected that health care professionals will balance the need for good image quality with the risks of radiation exposure on a case-by-case basis. It is not within the scope of this document to describe how these trade-offs should be resolved.

Recording of Anatomic Coverage by the Acquisition Device may or may not depend on attention and interaction by the Technologist.

The acquisition parameter constraints here have been selected with scans of the chest, abdomen and pelvis in mind.

### 3.2.2 Specification

| **Parameter** | **Actor** | **Specification** | **DICOM Tag** |
| --- | --- | --- | --- |
| Scan Duration for Thorax | Technologist | Shall achieve a table speed of at least 4cm per second, if table motion is necessary to cover the required anatomy. | Table Speed(0018,9309) |
| Scanogram | Technologist | Shall confirm on the scanogram the absence of artifact sources that could affect the planned volume acquisitions.  |  |
| Anatomic Coverage | Technologist | Shall ensure the tumors to be measured and additional required anatomic regions are fully covered. Shall, if multiple breath-holds are required, obtain image sets with sufficient overlap to avoid gaps within the required anatomic region(s), and shall ensure that each tumor lies wholly within a single breath-hold. | Anatomic Region Sequence(0008,2218) |
| Scan Plane (Image Orientation) | Technologist | Shall set Consistent with baseline. | Gantry/Detector Tilt (0018,1120) |
| Total Collimation Width | Technologist | Shall set to Greater than or equal to 16mm. | Total Collimation Width(0018,9307) |
| IEC Pitch | Technologist | Shall set to Less than 1.5. | Spiral Pitch Factor(0018,9311) |
| Tube Potential (kVp) | Technologist | Shall set Consistent with baseline (i.e. the same kVp setting if available, otherwise as similar as possible). | KVP (0018,0060) |
| Nominal Tomographic Section Thickness (T) | Technologist | Shall set to Less than or equal to 1.5mm. | Single Collimation Width(0018,9306) |
| Acquisition Field of View (FOV) | Technologist | Shall set Consistent with baseline. |  |
| Scan Capability | Acquisition Device | Shall be capable of performing scans with all the parameters set as described above in this table. |  |
| Image Header | Acquisition Device | Shall record actual Field of View, Scan Duration, Scan Plane, Total Collimation Width, Single Collimation Width, Scan Pitch, Tube Potential, Tube Current, Rotation Time, Exposure and Slice Width in the DICOM image header. |  |

## 3.3. Image Data Reconstruction

### 3.3.1 Discussion

Image reconstruction is modeled as a separate Activity in the QIBA Profile. Although it is closely related to image acquisition, and is usually performed on the Acquisition Device, reconstruction may be performed, or re-performed, separate from the acquisition. Many reconstruction parameters will be influenced or constrained by related acquisition parameters. This specification is the result of discussions to allow a degree of separation in their consideration without suggesting they are totally independent.

Many reconstruction parameters can have direct or indirect effects on identifying, segmenting and measuring tumors. To reduce this potential source of variance, all efforts should be made to have as many of the parameters as possible consistent with the baseline.

**Consistency** with the baseline implies a need for a method to record and communicate the baseline settings and make that information available at the time and place that subsequent reconstructions are performed. Although it is conceivable that the scanner could retrieve prior/baseline images and extract reconstruction parameters to encourage consistency, such interoperability mechanisms are not defined or mandated here and cannot be depended on to be present or used. Similarly, managing and forwarding the data files when multiple sites are involved may exceed the practical capabilities of the participating sites. Sites should be prepared to use manual methods instead.

**Spatial Resolution** quantifies the ability to resolve spatial details and scales the impact of partial volume effects. Lower spatial resolution can make it difficult to accurately determine the borders of tumors, and as a consequence, decreases the precision of volume measurements. Increased spatial resolution typically comes with an increase in noise which may degrade segmentation. If the spatial resolution is significantly different between the two timepoints, these impacts will change which can affect repeatability. So both balance and consistency is desirable. Maximum spatial resolution is mostly determined by the scanner geometry (which is not usually under user control) and the reconstruction kernel (over which the user has some choice).

Resolution is assessed (See 4.1) in terms of the f50 value of the modulation transfer function (MTF) measured in a scan of a resolution phantom (such as module 1 of the CTAP phantom from the American College of Radiology). An implication of using the ACR phantom is that the resolution is assessed at only one distance from the isocenter. Although spatial resolution may vary with distance from the isocenter and tumors can be expected at various distances from the isocenter, it is considered fair to assume that resolution does not degrade drastically relative to the acceptable range of the resolution specification here.

Note that the noise and resolution specifications (See 3.3.2) to “ensure that the protocol in use has been validated in phantoms“ are not asking the tech to scan phantoms before every patient, or to validate the protocol themselves, just that the site needs to have validated the protocols that the tech will be using and conformance with the protocol depends on the tech selecting those protocols.

**Voxel Noise Metrics** quantify the magnitude of the random variation in reconstructed CT numbers. Increased levels of noise can make it difficult to identify the boundary of tumors by humans and automated algorithms. If algorithms become uniformly more "noise tolerant", the maximum threshold may be raised. Decreased image noise is not always beneficial, if achieved through undesirable image manipulation (e.g. extreme amounts of image smoothing), or scanning technique (e.g. increases in radiation dose or decreases in resolution). The profile does not currently define a minimum threshold, although it could be introduced as a means of forcing a balance between the goal of noise reduction, and other priorities.

The preferred metric for voxel noise is the standard deviation of reconstructed CT numbers over a uniform region in a phantom. The use of standard deviation has limitations since it can vary with different reconstruction kernels, which will also impact the spatial resolution. While the Noise-Power Spectrum would be a more comprehensive metric, it is not practical to calculate (and interpret) at this time.

Voxel noise (pixel standard deviation in a region of interest) can be reduced by reconstructing images with greater thickness for a given mAs. It is not expected that the Voxel Noise be measured for each subject scan, but rather the Acquisition Device and Reconstruction Software be qualified for the expected acquisition and reconstruction parameters.

**Reconstruction Field of View** affects reconstructed pixel size because the fixed image matrix size of most reconstruction algorithms is 512x512. If it is necessary to expand the field of view to encompass more anatomy, the resulting larger pixels may be insufficient to achieve the claim. A targeted reconstruction with a smaller field of view may be necessary, but a reconstruction with that field of view would need to be performed for every time point. Pixel Size directly affects voxel size along the subject x-axis and y-axis. Smaller voxels are preferable to reduce partial volume effects and provide higher measurement precision. Pixel size in each dimension is not the same as spatial resolution in each dimension. The spatial resolution of the reconstructed image depends on a number of additional factors including a strong dependence on the reconstruction kernel.

**Reconstructed Image Thickness** is the nominal width of the reconstructed image along the z-axis (reconstructed image thickness) since the thickness is not technically the same at the middle and at the edges.

**Reconstructed Image Interval** is the distance between two consecutive reconstructed images. An interval that results in discontiguous data is unacceptable as it may “truncate” the spatial extent of the tumor, degrade the identification of tumor boundaries, confound the precision of measurement for total tumor volumes, etc. Decisions about overlap (having an interval that is less than the nominal reconstructed slice thickness) need to consider the technical requirements of the clinical trial, including effects on measurement, throughput, image analysis time, and storage requirements.

Reconstructing datasets with **overlap** will increase the number of images and may slow down throughput, increase reading time and increase storage requirements. For multi-detector row CT (MDCT) scanners, creating overlapping image data sets has NO effect on radiation exposure; this is true because multiple reconstructions having different kernel, slice thickness and intervals can be reconstructed from the same acquisition (raw projection data) and therefore no additional radiation exposure is needed.

**Reconstruction Kernel** Characteristics influence the texture and the appearance of tumors in the reconstructed images, which may influence measurements. A softer kernel can reduce noise at the expense of spatial resolution. An enhancing kernel can improve resolving power at the expense of increased noise. The characteristics of different tissues (e.g. lung) may call for the use of different kernels, and implementers are encouraged to use kernels suitable for the anatomic region and tissue imaged. The use of multiple kernels in a single study is not prohibited by the specification below, but any given tumor must be measured on images reconstructed using consistent kernels at each time point.

**Algorithm Type** may influence the texture and the appearance of tumors in the reconstructed images, which may influence measurements. However the effect of different <TODO resolve seeming conflict between this sentence and the previous> iterative reconstruction algorithms on the variability of noise, margin conspicuity, and segmentation is considered to be insufficient to impact the claimed performance.

Although the profile requires that the type of algorithm (Model-based iterative vs Statistical iterative vs FBP) may not be different between the baseline and subsequent scans, it is permitted to use different reconstruction software (e.g. different manufacturers) and still achieve the performance claim. (Mention the rationale that having the constraints on resolution and noise, then the reconstruction impact is probably not that significant) (If iterative recon has blurred partial solid into full solid, you have probably violated the resolution constraint) (Should we plan a groundwork project to confirm this, or build checking it into our field test?)

The **stability of HU** between time points and its effect on volume measurements is not fully understood as of the writing of this version of the Profile.

### 3.3.2 Specification

**Note:** The Radiologist is responsible for the protocol parameters and validation, although they may be executed by a medical physicist or other qualified staff (such as vendor service or specialists).

| **Parameter** | **Actor** | **Specification** |
| --- | --- | --- |
| Reconstruction Protocol | Technologist | Shall select a protocol that has been prepared and validated for this purpose.Shall report if any parameters are modified beyond these specifications. |
| ReconstructionField of View | Technologist | Shall ensure the Field of View spans at least the full extent of the thoracic and abdominal cavity, but not substantially greater than that, and is consistent with baseline. |
| In-plane Spatial Resolution | Radiologist | Shall ensure that the protocol has been validated to achieve an f50 value that is between 0.3 mm-1 and 0.7 mm-1, and is within 0.2 mm-1 of the baseline.See 4.1. Assessment Procedure: In-plane Spatial Resolution |
| Voxel Noise  | Radiologist | Shall ensure that the protocol has been validated to achieve: * a standard deviation that is < 50HU and consistent with the baseline scan within 5HU.
* an favg that is between 0.25 mm-1 and 0.35 mm-1, and is within 0.05 mm-1 of the baseline.

See 4.2. Assessment Procedure: Voxel Noise |
| Reconstructed Image Thickness | Radiologist | Shall set to between 1.0mm and 2.5mm (inclusive) and consistent (i.e. within 0.5mm) with baseline. |
| Reconstructed Image Interval | Radiologist | Shall set to less than or equal to the Reconstructed Image Thickness (i.e. no gap, may have overlap) and consistent with baseline. |
| Reconstruction Algorithm Type | Radiologist | Shall set to be consistent with baseline (i.e. Filtered Back-Projection used for both, or Model-based Iterative used for both, or Statistical Iterative used for both). |
| Reconstruction Kernel Characteristics | Radiologist | Shall set Consistent with baseline (i.e. the same kernel if available, otherwise the kernel most closely matching the kernel response of the baseline).  |
| Reconstruction Capability | Reconstruction Software | Shall be capable of performing reconstructions with all the parameters set as described above and producing images as described above. |
| Image Header | Reconstruction Software | Shall record actual Spatial Resolution, Noise, Pixel Spacing, Reconstruction Interval, Reconstruction Overlap, Reconstruction Kernel Characteristics, as well as the model-specific Reconstruction Software parameters utilized to achieve compliance with these metrics in the image header. |

## 3.4. Image QA

### 3.4.1 Discussion

This Image QA activity represents the portion of QA performed between image generation and analysis where characteristics of the content of the image are checked for compliance with the profile. The Image QA details listed here are the ones QIBA has chosen to highlight in relation to achieving the Profile claim. It is expected that sites will perform many other QA procedures as part of good imaging practices.

The Radiologist is identified here as ultimately responsible for this activity; however sites may find it beneficial for technologists to review these details at the time of imaging and identify cases which might require repeating acquisition and/or reconstruction to address issues with patient motion or artifacts.

Similarly, some or all of these checks may be performed at reporting time and as a result some or all of the tumor measurements may then be identified as not falling within the performance Claim of the Profile.

**Patient motion artifacts** can manifest in a variety of ways, such as a perceptible tram tracking appearance of the bronchioles or blurring of the lung architectural contours with lung windows.

**Dense object artifacts** (both internal and external to the patient) can variably degrade the ability to assess tumor boundaries as discussed in 3.1.4.1, resulting in poor change measures and repeatability.

**Clinical conditions** can also degrade the ability to assess tumor boundaries, or influence the structure of the tumor itself. For example, atelectasis, pleural effusion, pneumonia and/or pneumothorax can result in architectural changes to the lung surrounding a nodule. Necrosis may complicate decisions on the tumor extent.

**Tumor Size** can affect the accuracy of measurements. Both theoretical considerations and the groundwork projects done by QIBA indicate that for tumors that are small, errors in measurement represent a greater percentage of the measured size. For tumors that are smaller than the limits defined in this profile, please see the profile produced by the QIBA Small Nodule group for more information on imaging recommendations and performance claims. For tumors that are extremely large, the limitations on measurement are based less on imaging physics and more on anatomy. Such tumors are likely to cross anatomical boundaries and abut structures that make consistent segmentation difficult.

**Tumor Margin Conspicuity** refers to the clarity with which the boundary of the tumor can be discerned from the surroundings. Conspicuity can directly impact the ability to segment the tumor to properly determine its volume. Conspicuity problems can derive from poor contrast enhancement, from the inherent texture, homogeneity or structure of the tumor, or from attachment of the tumor to other structures.

**Tumor Measurability** is a general evaluation that is essentially left to the judgement of the radiologist, and it is their responsibility to oversee segmentation and disqualify tumors with poor measurability or inconsistent segmentation between the two timepoints. If the tumor has varying margin conspicuity on different slices, or is conspicuous but has complex geometry, or the segmentation software is visibly failing, or the background didn't respond to contrast the same way in the two time points, the radiologist may disqualify the tumor. Conversely, if the tumor is attached to another structure but the radiologist is confident they can get consistent segmentation over the two timepoints, they may allow a tumor that would be otherwise disqualified.

**Tumor Shape** is not explicitly called out as a specification parameter. No specific tumor shapes are considered a priori unsuitable for measurement. Although groundwork has shown that consistent measurements are more readily achieved with simple shapes than with complex shapes (such as spiculated tumors), the parameters for tumor size, tumor margin conspicuity and tumor measurability are felt to be sufficient. Moreover, complex shapes are even more difficult to assess accurately using simple linear measurements, increasing the relative added value of volumetry.

Keep in mind that this Profile is “lesion-oriented”. If one tumor in a study is excluded from the Profile Claim because the tumor does not comply with the specifications in this section, that does not affect other tumors in the same study which do comply with these specifications at both time points. Further, if a future study results in the excluded tumor being compliant at two time points, then the claim holds across those two time points.

While the radiologist is responsible for confirming case compliance with the Image QA specifications in Section 3.4.2, it is left to individual sites to determine the best approach in their work environment for capturing this audit data. Possible approaches include the use of a QIBA worksheet that captures this information, or asking the radiologist to dictate each parameter into the clinical report (e.g. the scan is free of motion or dense object artifacts, contrast enhancement is consistent with baseline, the tumor margins are sufficiently conspicuous").

### 3.4.2 Specification

The Radiologist shall ensure that the following specifications have been evaluated for each tumor being measured.

| **Parameter** | **Actor** | **Specification** |
| --- | --- | --- |
| Patient Motion Artifacts | Radiologist | Shall confirm the images containing the tumor are free from artifact due to patient motion. |
| Dense Object Artifacts | Radiologist | Shall confirm the images containing the tumor are free from artifact due to dense objects, materials or anatomic positioning.  |
| Clinical Conditions | Radiologist | Shall confirm that there are no clinical conditions affecting the measurability of the tumor.  |
| Tumor Size | Radiologist | Shall confirm (now or during measurement) that tumor longest in-plane diameter is between 10 mm and 100 mm. (For a spherical tumor this would roughly correspond to a volume between 0.5 cm3 and 524 cm3.) |
| Tumor Margin Conspicuity | Radiologist | Shall confirm the tumor margins are sufficiently conspicuous and unattached to other structures of equal density to distinguish the volume of the tumor. |
| Contrast Enhancement | Radiologist | Shall confirm that the phase of enhancement and degree of enhancement of appropriate reference structures (vascular or tissue) are consistent with baseline.  |
| Tumor Measurability | Radiologist | Shall disqualify any tumor they feel might reasonably degrade the consistency and accuracy of the measurement.Conversely, if artifacts or attachments are present but the radiologist is confident and prepared to edit the contour to eliminate the impact, then the tumor need not be judged non-compliant with the Profile. |
| Consistency with Baseline | Radiologist | Shall confirm that the tumor is similar in both timepoints in terms of all the above parameters. |

## 3.5. Image Analysis

### 3.5.1 Discussion

This Profile characterizes each designated tumor by its volume change relative to prior image sets.

This is typically done by determining the boundary of the tumor (referred to as segmentation), computing the volume of the segmented tumor and calculating the difference of the tumor volume in the current scan and in the baseline scan.

Volume Calculation values from a segmentation may or may not correspond to the total of all the segmented voxels. The algorithm may consider partial volumes, do surface smoothing, tumor or organ modeling, or interpolation of user sculpting of the volume. The algorithm may also pre-process the images prior to segmentation.

Segmentation may be performed automatically by a software algorithm, manually by a human observer, or semi-automatically by an algorithm with human guidance/intervention, for example to identify a starting seed point, stroke, or region, or to edit boundaries.

If a human observer participates in the segmentation, either by determining while looking at the images the proper settings for an automated process, or by manually editing boundaries, the settings for conversion of density into display levels (window and level) should either be fixed during the segmentation process or documented so that observers can apply consistent display settings at future scans (or a different observer for the same scan, if multiple readers will read each scan, as for a clinical trial).

**Tumor Volume Accuracy** can affect the variability of Tumor Volume Change results. The volume accuracy is assessed to confirm that volume is being computed correctly and confirm there is a reasonable lack of bias at individual timepoints.

**Tumor Volume Change Variability**, which is the focus of the Profile Claim, is a key performance parameter for this biomarker. The 30% target is a conservative threshold of measurement variation (the 30% change in the claim is the outside of 95% confidence interval of 15% of measurement variability when sample size is 40 or more). Based on a survey of clinical studies (See Appendix B.2) the 30% target is considered to be reasonable and achievable. In Table B.1, the range between the minimum and maximum values in the 95% CI of the measurement difference column is mostly within +/- 15%.

Methods that calculate volume changes directly without calculating volumes at individual time points are acceptable so long as the results are compliant with the specifications set out by this Profile.

The Image Analysis Tool should be prepared to process both the current data and previous data at the same time and support matching up the appearance of each tumor in both data sets in order to derive volume change values. Although it is conceivable that they could be processed separately and the results of prior processing could be imported and a method of automated tagging and matching of the tumors could be implemented, such interoperability mechanisms are not defined or mandated here and cannot be depended on to be present or used.

**Reading Paradigms** (such as the “sequential locked” paradigm described here) can reduce variability from inconsistent judgments (such as where to separate an attached tumor) but also have the potential to introduce subconscious biases. The current edition of the profile does not prohibit the Image Analysis Tool from displaying the actual volume value from the previous timepoint since that might unnecessarily disqualify existing products. If it is determined to be the source of problems, it might be prohibited in future editions. Also, note that while the Image Analysis Tool is required to be capable of displaying the image from the previous timepoint, the radiologist is not required to look at it for every case. It is up to their judgment when to use that capability.

Storing segmentations and measurement results that can be loaded by an Image Analysis Tool analyzing data collected at a later date is certainly a useful practice as it can save time and cost. For this to happen reliably, the stored format must be compatible and the data must be stored and conveyed. Although DICOM Segmentation objects are appropriate to store tumor segmentations, and DICOM SR objects are appropriate to store measurement results, these standards are not yet widely enough deployed to make support for them mandatory in this Profile. Similarly, conveying the segmentations and measurements from baseline (and other time points prior to the current exam) is not done consistently enough to mandate that it happen and to require their import into the Image Analysis Tool. Managing and forwarding the data files may exceed the practical capabilities of the participating sites.

Image analysis can be performed on any equipment that complies with the specifications set out in this Profile. However, we strongly encourage performing all analysis for an individual subject on the same platform (manufacturer, model and version) which we expect will further reduce variation.

Medical Devices such as the Image Analysis Tool are typically made up of multiple components (the hardware, the operating system, the application software, and various function libraries within those). Changes in any of the components can affect the behavior of the device. In this specification, the “device version” should reflect the total set of components and any changes to components should result in a change in the recorded device version. This device version may thus be different than the product release version that appears in manufacturer documentation.

For analysis methods that involve an operator (e.g. to draw or edit boundaries, set seed points or adjust parameters), the operator is effectively a component of the system, with an impact on the reproducibility of the measurements, and it is important to record the operator’s identify as well. Fully automated analysis software removes that source of variation; although even then, since a human is generally responsible for the final results, they retain the power to approve or reject measurements so their identity should be recorded.

The Tumor Volume Change performance specification below includes the operator performance and is intended to be evaluated based on a typical operator (i.e. without extraordinary training or ability). This should be kept in mind by manufacturers measuring the performance of their tools and sites validating the performance of their installation. Although the performance of some methods may depend on the judgment and skill of the operator, it is beyond this Profile to specify the qualifications or experience of the operator.

Determination of which tumors should be measured is out of scope for this Profile. Such determination may be specified within a protocol or specified by formal response criteria standards, or may be determined by clinical requirements. Tumors to be measured may be designated by the oncologist or clinical investigator, by a radiologist at a clinical site, by a reader at a central reading facility, or they may be designated automatically by a software analysis tool.

**Confidence Interval of Result** provides a range of plausible values for the change in tumor volume. The 95% confidence interval (CI) can be interpreted as follows: If change in a tumor's volume is measured repeatedly and the 95% CI constructed for each measurement, then 95% of those CIs would contain the true volume of the tumor.

A reference implementation of a calculator that uses the specified equation is available at the following location: <http://www.accumetra.com/NoduleCalculator.html>

<Note that 9% wCV is assumed constant (based on same scanner/SW/radiologist for both timepoints) over tumor sizes but that might not be strictly true>

**Recording** various details can be helpful when auditing the performance of the biomarker and the site using it. For example, it is helpful for the system to record the set-up and configuration parameters used, or to be capable of recording the tumor segmentation as a DICOM Segmentation. Systems based on models should be capable of recording the model and parameters.

It is up to products that do not use contours to propose a method for verification by the radiologist.

### 3.5.2 Specification

| **Parameter** | **Actor** | **Specification** |
| --- | --- | --- |
| Multiple Tumors | Image Analysis Tool | Shall allow multiple tumors to be measured.Shall either correlate each measured tumor across time points or support the radiologist to unambiguously correlate them. |
| Reading Paradigm | Image Analysis Tool | Shall be able to present the reader with both timepoints side-by-side for comparison when processing the second timepoint. |
| Tumor Volume Computation | Image Analysis Tool | Shall be validated to compute tumor volume with accuracy within 3 % of the true volume.See 4.3 Assessment Procedure: Tumor Volume Computation. |
| Tumor VolumeChange Repeatability | Image Analysis Tool | Shall be validated to achieve tumor volume change repeatability with: * an overall repeatability coefficient of less than 16%.
* a small subgroup repeatability coefficient of less than 21%
* a large subgroup repeatability coefficient of less than 21%

See 4.4. Assessment Procedure: Tumor Volume Change Repeatability.  |
| Radiologist | Shall, if operator interaction is required by the Image Analysis Tool to perform measurements, be validated to achieve tumor volume change repeatability with:* an overall repeatability coefficient of less than 16%.
* a small subgroup repeatability coefficient of less than 21%
* a large subgroup repeatability coefficient of less than 21%

See 4.4. Assessment Procedure: Tumor Volume Change Repeatability (Image Analysis Tool). |
| Tumor Volume Bias& Linearity | Image Analysis Tool | Shall be validated to achieve:* an overall tumor volume %bias of less than shown in Table 3.5.2-2 (below)
* a tumor volume %bias for each shape subgroup (spherical, ovoid, lobulated) of less than shown in Table 3.5.2-2 (below)
* slope ( between 0.98 and 1.02

Values are taken from Table 3.5.2-2 based on the overall repeatability coefficient achieved by the Image Analysis Tool using the assessment procedure in 4.4. See 4.5 Assessment Procedure: Tumor Volume Bias and Linearity. |
| ResultVerification | Radiologist | Shall review & approve margin contours produced by the tool. |
| Confidence Interval of Result | Image Analysis Tool | Shall calculate and make available to the operator the 95% confidence interval for tumor volume change based on the equation:Where  are the volume measurement at timepoint 2 and 1 is the within-nodule coefficient of variation which is 9% |
| Recording | Image Analysis Tool | Shall record the percentage volume change relative to baseline for each tumor. Shall record the calculated confidence interval for each tumor. TODO REVIEWShall record the image analysis tool version.  |

**Table 3.5.2-2:
Allowable Tumor Volume %Bias based on Repeatability Coefficient**

|  |  |  |
| --- | --- | --- |
| **Repeatability Coefficient p** | **Allowable Overall %Bias** | **Allowable Shape Subgroup %Bias** |
| 5% | <7.2% |  |
| 6% | <7.0% |  |
| 7% | <6.8% |  |
| 8% | <6.5% |  |
| 9% | <6.3% |  |
| 10% | <5.9% |  |
| 11% | <5.5% |  |
| 12% | <5.1% |  |
| 13% | <4.5% |  |
| 14% | <3.8% |  |
| 15% | <2.8% |  |
| 16% | <1.2% |  |

# 4. Assessment Procedures

To conform to this Profile, participating staff and equipment (“Actors”) shall support each of the activities assigned to them in Table 1.

For each activity, the conformance requirements checklist (sometimes referred to as the “shall language”) for each Actor are documented in Section 3.

Although most of the requirements described in Section 3 are feature-oriented and/or conformance can be assessed by direct observation, some of the requirements are performance-oriented. The following sub-sections elaborate on the meaning of performance-oriented requirements and how they are intended to be correctly assessed.

Formal claims of conformance by the organization responsible for an Actor shall be in the form of a published QIBA Conformance Statement. Manufacturers publishing a QIBA Conformance Statement shall provide a set of “Model-specific Parameters” (as shown in Appendix D) describing how their product was configured to achieve conformance. Manufacturers shall also provide access or describe the characteristics of the test set used for compliance testing.

## 4.1. Assessment Procedure: In-plane Spatial Resolution

This procedure can be used by a manufacturer or an imaging site to assess the In-plane Spatial Resolution of reconstructed images. Resolution is assessed in terms of the f50 value (in mm-1) of the modulation transfer function (MTF). Loosely speaking, the MTF represents the blur of an infinitely small feature of interest, f50 represents the spatial frequency at which the contrast of the feature has decreased by 50%, and the inverse of the f50 value represents the size of a feature that would be degraded 50%. So for an f50 value of 0.4 mm-1, features that are 2.5mm (or smaller) would have their contrast degraded by 50% (or more).

The assessor shall first warm up the scanner’s x-ray tube and perform calibration scans (often called air-calibration scans) according to scanner manufacturer recommendations. The assessor shall scan a spatial resolution phantom, such as the ACR CT Accreditation Program (CTAP) Phantom’s module 1, which has a series of HU-value cylindrical inserts including one with soft-tissue equivalence. The acquisition protocol and reconstruction parameters shall conform to this Profile (See Section 3.2.2 and 3.3.2). The same protocol and parameters shall be used when performing the assessments in 4.1 and 4.2. I.e., the noise level during resolution assessment should correspond to that measured during noise assessment.

The phantom shall be positioned with the center of the phantom at isocenter and properly aligned along the z-axis as described in the ACR CTAP documentation about alignment of the beads.

When the scan is performed, the assessor shall generate an MTF curve, measured as an average of the MTF in the x-y plane along the edge of a target soft-tissue equivalent insert using AAPM TG233 methodology as implemented in manufacturer analysis software, AAPM TG233 software or equivalent.

The assessor shall then determine and record the f50 value, defined as the spatial frequency (in mm-1 units) corresponding to 0.5 MTF on the MTF curve.

The procedure described above is provided as a reference method. This reference method and the method used by the scanner manufacturer for FDA submission of MTF values are accepted methods for this assessment procedure. Note that for iterative reconstruction, the manufacturer may have specific test methodologies appropriate for the given algorithm.

Sites may submit to QIBA a proposed alternative method and evidence that the results produced by the proposed method are equivalent to this reference method or to the manufacturer method. Upon review and approval by QIBA, the alternative method will also become an accepted assessment procedure in this Profile.

The test procedure described here may be applied to both conventional filtered backprojection reconstruction methods and iterative reconstruction methods. It is expected that the performance of tumor segmentation and volume estimation should be robust across FBP and iterative methods.

The assessor is recommended to repeat the assessment with the center of the phantom positioned 20cm away from isocenter and record the resolution results as part of the assessment record. The easiest way to achieve this may be by adjusting the table height after completing the first assessment scan.

## 4.2. Assessment Procedure: Voxel Noise

This procedure can be used by a manufacturer or an imaging site to assess the voxel noise of reconstructed images. Voxel noise is assessed in terms of the standard deviation of pixel values (which gives a sense of the noise magnitude) when imaging a material with uniform density.

The assessor shall first warm up the scanner’s x-ray tube and perform calibration scans (often called air-calibration scans) according to scanner manufacturer recommendations. The assessor shall then scan a phantom of uniform density, such as the ACR CT Accreditation Program (CTAP) Phantom’s module 3, which is a 20 cm diameter cylinder of water equivalent material. The phantom shall be placed at the isocenter of the scanner. The acquisition protocol and reconstruction parameters shall be compliant with this Profile (See Section 3.2.2 and 3.3.2). The same protocol and parameters shall be used when performing the assessments in 4.1 and 4.2.

When the scan is performed, the assessor shall select a single representative slice from the uniformity portion of the phantom.

An approximately circular region of interest (ROI) of at least 400 mm2 shall be placed near the center of the phantom. The assessor shall record the values reported for the ROI mean and standard deviation.

The procedure described above is provided as a reference method. Sites may submit to QIBA a proposed alternative method (such as using the water phantom portion of a manufacturer’s QA phantom) and evidence that the results produced by the proposed method are equivalent to this reference method or manufacturer methodology. Upon review and approval by QIBA, the alternative method will also become an accepted assessment procedure in this Profile.

The test procedure described here is based on the use of both conventional filtered backprojection and iterative reconstruction methods; however, care must be taken when voxel noise alone is used to characterize reconstruction methods as noise voxel is a limited representation of image noise when noise texture is varied.

For a discussion of considerations when performing this procedure, see:

1. McCollough, et al, “The phantom portion of the ACR CT accreditation program: Practical tips, artifact examples, and pitfalls to avoid”, Medical Physics, Vol. 31, No. 9, September 2004 <http://www.aapm.org/meetings/05am/pdf/18-4146-57655-316.pdf>
2. The ACR CT QC manual; information about obtaining this manual is available at <http://www.acr.org/Education/Education-Catalog/Products/8336734>
3. Wilson JM, Christianson OI,Richard S, Samei E. A methodology for image quality evaluation of advanced CT systems. *Medical Physics* 40(3): 031908-01-09, 2013.
4. Chen B, Samei E. Development of a phantom-based methodology for the assessment of quantification performance in CT. SPIE International Symposium on Medical Imaging, Orlando, FA, February 2013, *Proc. SPIE* *Medical Imaging* 8668: 86681E, 2013.
5. Chen B, Christianson O, Wilson J, Samei E. Assessment of volumetric noise and resolution performance for linear and nonlinear CT reconstruction methods. *Medical Physics* 41, 071909, 2014.
6. Christianson O, Chen J, Yang Z, Saiprasad G, Dima A, Filliben J, Peskin A, Trimble C, Siegel E, Samei E. An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations. *Radiology* 275(3): 725-734, 2015.

## 4.3. Assessment Procedure: Tumor Volume Computation

This procedure can be used by a manufacturer or an imaging site to assess whether an Image Analysis Tool computes the volume of a single tumor correctly. Accuracy is assessed in terms of the percentage error when segmenting and calculating the volume of a tumor with known truth.

The assessor shall obtain the test files in DICOM format from the QIDW. They can be found by searching for the CT volumetry digital reference object (DRO) DICOM image set. The test files represent a digital test object with z-axis resolution of 1.5mm. A test nodule with -10 HU radio-density is placed within a flat -1000 HU region of the phantom to make the segmentation intentionally easy since the test is not intended to stress the segmentation tool but to instead evaluate any bias in the volume computation after the lesion is segmented.

The assessor shall use the Image Analysis Tool to segment and calculate the volume of the single tumor present in the test images.

The assessor shall record the percentage difference between the reported volume and the true value in the description of the test files on QIDW.

## 4.4. Assessment Procedure: Tumor Volume Change Repeatability

This procedure can be used by a manufacturer or an imaging site to assess the repeatability with which the volume of a single tumor is measured. Repeatability is assessed in terms of the repeatability coefficient when segmenting and calculating the volume of a tumor with known truth. The procedure may be used to assess an Image Analysis Tool or a Radiologist operating such a tool.

The assessment procedure has the following steps:

* Obtain a designated test image set (see 4.4.1).
* Determine the volume change for designated tumors (see 4.4.2).
* Calculate statistical metrics of performance (see 4.4.3).

Note that tumor detection is not evaluated by this procedure since the locations of the target lesions are provided.

### 4.4.1 obtain test image set

The test image set consists of multiple target tumors in the lung in multiple subjects which is representative of the stated scope of the Profile.

The assessor shall obtain the test files in DICOM format.

Lung tumor data is obtained from the Cancer Imaging Archive by searching for the Coffee-break subset of the RIDER Lung CT Dataset at (https://public.cancerimagingarchive.net/ncia/login.jsf).

<<We will not publish the Technically Confirmed Profile without replacing this dataset, so we don't have the problem of algorithms training on the test set. Consider adding at least a couple of single timepoint liver studies to segment multiple times. And/or get Ehsan to generate some known truth images. >>

The test files represent 31 cases, with two time points per case, each with one target tumor to segment. The target tumor is identified in terms of its x/y/z coordinates in the dataset. The list of target tumors and coordinates are provided in file: (???)

Future editions of the Profile may address a larger number of body parts (e.g., metastases in the mediastinum, liver, adrenal glands, neck, retroperitoneum, pelvis, etc. described in Appendix B.3) by including such tumors in the test data, and may test boundary condition performance by including test data that is marginally conformant (e.g. maximum permitted slice thickness, maximum permitted noise, etc.) to confirm conformant performance is still achieved.

The target tumors have been selected to be measureable (as defined in the Profile) and have a range of volumes, shapes and types to be representative of the scope of the Profile.

The test image set has been acquired according to the requirements of this Profile (e.g. patient handling, acquisition protocol, reconstruction).

If the algorithm has been developed using the specified test files, that is unfortunate and shall be reported by the assessor.

### 4.4.2 determine volume change

The assessor shall segment each target tumor at each timepoint as described in the Image Analysis Activity (See 3.5). The assessor is permitted to edit the tumor segmentation or seed point if that is part of the normal operation of the tool. If segmentation edits are performed, results shall be reported both with and without editing.

When evaluating an Image Analysis Tool, a single reader shall be used for this entire assessment procedure.

When evaluating a Radiologist, a single tool shall be used for this entire assessment procedure.

The assessor shall calculate the volume (Y) of each target tumor at time point 1 (denoted Y*i*1) and at time point 2 (Y*i*2) where *i* denotes the *i*-th target tumor.

The assessor shall calculate the resulting % volume change (d) for each target tumor as

.

### 4.4.3 calculate statistical metrics of performance

The assessor shall calculate the within-subject Coefficient of Variation (wCV), where N=31 and

The assessor shall estimate the Repeatability Coefficient (RC) as

The assessor shall convert the Repeatability Coefficient (RC) estimate to a percentage as

.

The assessor shall divide the target tumors into a small subgroup (containing the 15 target tumors with the smallest measured volumes) and a large subgroup (containing the 16 tumors with the largest measured volumes). The assessor shall repeat the above calculations on both subgroups to estimate a small subgroup repeatability coefficient and a large subgroup repeatability coefficient.

The assessor is recommended to also compute Bland-Altman plots of the volume estimates as part of the assessment record.

For further discussion/rationale, see Annex E.2 Considerations for Performance Assessment of Tumor Volume Change.

## 4.5. Assessment Procedure: Tumor Volume Bias and Linearity

This procedure can be used by a manufacturer or an imaging site to assess the bias and linearity with which the volume of a single tumor is measured. Bias is assessed in terms of the percentage population bias when segmenting and calculating the volume of a number of tumors with known truth. Linearity is assessed in terms of the slope of an OLS regression fit to the volume data.

### 4.5.1 obtain test image set

The test image set consists of scans from two different scanners of an anthropomorphic ("Lungman") phantom with multiple synthetic target tumors of different shapes and sizes in the lung.

The assessor shall obtain the test files in DICOM format.

The data is obtained from the Cancer Imaging Archive by searching for the <<Nick will clarify the location of the FDA Lungman N1 data>> subset of the RIDER Lung CT Dataset at (https://public.cancerimagingarchive.net/ncia/login.jsf).

The test files represent 3 repeated scans of the FDA Lungman N1 phantom on each of 2 CT scanners. The phantom contains 7 synthetic tumors, each with a different combination of shape and diameter (see Table 4.5.1-1). The target tumors are identified in terms of their x/y/z coordinates in each of the 6 scans. The list of target tumors and coordinates are provided in file: (???)

Table 4.5.1-1: Phantom Target Tumor Characteristics

|  |  |  |
| --- | --- | --- |
| **Shape** | **Nominal Diameter** | **Nominal Density** |
| Spherical | 10 mm20 mm40 mm | +100 HU |
| Ovoid | 10 mm20 mm | +100 HU |
| Lobulated | 10 mm20 mm | +100 HU |

The target tumors have been placed to be measureable (as defined in the Profile) and have a range of volumes and shapes to be representative of the scope of the Profile.

The test image set has been acquired according to the requirements of this Profile (e.g. patient handling, acquisition protocol, reconstruction). See Table 4.5.1-2.

Table 4.5.1-2: Test Image Set Acquisition and Reconstruction Parameters

|  |  |
| --- | --- |
| **Scanner** | **Key Parameters** |
| Philips 16(Mx8000 IDT) | KVp: 120Pitch: 1.2Collimation: 16x1.5Exposure: 100 mAsSlice Thickness: 2 mmIncrement: 1 mmFilter: MediumRepeat Scans: 3 |
| Siemens 64 | KVp: 120Pitch: 1.2Collimation: 64x0.6Exposure: 100 mAsSlice Thickness: 1.5 mmIncrement: 1.5 mmFilter: MediumRepeat Scans: 3 |

### 4.5.2 determine volume change

The assessor shall segment each of 42 target tumors (7 tumors in 3 scans for each of 2 scanners) as described in the Image Analysis Activity (See 3.5).

The assessor is permitted to edit the tumor segmentation or seed point if that is part of the normal operation of the tool. If segmentation edits are performed, results shall be reported both with and without editing.

When evaluating an Image Analysis Tool, a single reader shall be used for this entire assessment procedure.

When evaluating a Radiologist, a single tool shall be used for this entire assessment procedure.

The assessor shall calculate the volume (Y) of each target tumor (denoted Y*i*) where *i* denotes the *i*-th target tumor.

### 4.5.3 calculate statistical metrics of performance

The natural log of the true volumes (Xi) of each target tumor are known and are provided in the dataset.

The assessor shall calculate the individual bias (*bi*) of the measurement of each target tumor as

The assessor shall estimate the population bias over the N target tumors as

The assessor shall convert to a percentage bias estimate as

 =

The assessor shall fit an ordinal least squares (OLS) regression of the on and shall estimate the slope ).

The assessor shall divide the target tumors into three subgroups (containing the spherical, ovoid and lobulated target tumors respectively). The assessor shall repeat the percentage population bias calculation on each subgroup to estimate a spherial subgroup percentage bias, an ovoid subgroup percentage bias and a lobulated subgroup percentage bias.

The assessor is recommended to also plot the volume estimate ( versus ) and the OLS regression curve of the volume estimates as part of the assessment record.

## 4.X. Assessment Procedure: Image Acquisition Site Performance

Note: The procedure in this section is currently only a proposal.

A more detailed procedure will be provided in the future.

Until then, there is no approved way to claim conformance to this performance requirement.

Site performance can be assessed with the following procedure:

* Validate image acquisition (see 4.X.1).
* Generate a test image set (see 4.X.2).
* Assess Tumor Volume Change Variability (see 4.1.2, 4.1.3 above).
* Compare against the Tumor Volume Change Variability performance level specified in 3.5.2.

This procedure can be used by an imaging site to evaluate the performance of each of the Actors and Activities in use. In principle, the final result represents an assessment of the combined performance of all the Actors and Activities at the site.

The procedure presumes that the Actors being used by the site are capable of meeting the requirements described in Section 3 of this document; however it is not a pre-requisite that those Actors have published QIBA Conformance Statements (although that would be both useful and encouraging).

Discussion:

Duke is working on a “platform” that includes a phantom and an analysis tool that may inform the future contents of this section.

Sites that carry out this procedure should really record the parameters they used and document them in something similar to a Conformance Statement. This would be a useful QA record and could be submitted to clinical trials looking for QIBA compliant test sites.

Are there other criteria that should be worked into this procedure?

Typically clinical sites are selected due to their competence in oncology and access to a sufficiently large patient population under consideration. For imaging it is important to consider the availability of:

 - appropriate imaging equipment and quality control processes,

 - appropriate injector equipment and contrast media,

 - experienced CT Technologists for the imaging procedure, and

 - processes that assure imaging Profile compliant image generation at the correct point in time.

A clinical trial might specify “A calibration and QA program shall be designed consistent with the goals of the clinical trial. This program shall include (a) elements to verify that sites are performing correctly, and (b) elements to verify that sites’ CT scanner(s) is (are) performing within specified calibration values. These may involve additional phantom testing that address issues relating to both radiation dose and image quality (which may include issues relating to water calibration, uniformity, noise, spatial resolution -in the axial plane-, reconstructed slice thickness z-axis resolution, contrast scale, CT number calibration and others). This phantom testing may be done in additional to the QA program defined by the device manufacturer as it evaluates performance that is specific to the goals of the clinical trial.”

### 4.X.1 Acquisition Validation

Review patient handling procedures for compliance with Section 3.1

Establish acquisition protocols and reconstruction settings on the Acquisition Device compliant with Section 3.2 and Section 3.3. If a QIBA Conformance Statement is available from the Acquisition Device manufacturer, it may provide parameters useful for this step.

Acquire images of a 20cm water phantom, reconstruct and confirm performance requirements in Section 3.3.2 are met.

Discussion:

UCLA may have more detailed and more complete procedures to recommend for this section.

### 4.X.2 Test Image Set

Locally acquire a test image set using the protocols established and tested in Section 4.X.1.

The test image set should conform to the characteristics described in Section 4.X.1.

Discussion:

It is highly likely that due to practical constraints the test image set prepared at an individual site would be much less comprehensive than the test image sets prepared by QIBA. Further consideration of what a more limited but still useful test image set would look like.

#

# Closed Issues:

The following issues have been considered closed by the technical committee. They are provided here to forestall discussion of issues that have already been raised and resolved, and to provide a record of the rationale behind the resolution. It will be removed during publication of the Technically Confirmed Draft.

|  |  |
| --- | --- |
| **1** | **Q. Is the claim appropriate/supported by the profile details, published literature, and QIBA groundwork? Is it stated in clear and statistically appropriate terms?**A. Basically, yes.Claim reworded to be clear and statistically appropriate. The concept of “levels of confidence” has been introduced (See separate documents and process). Claim seems to be appropriate for the “Reviewed” level of confidence.In terms of anatomy, it is recognized that the acquisition protocols and processing will not be appropriate for all types of tumors in all parts of the body, however it is felt that the conspicuity requirements will make it clear to users of the profile which anatomy is not included. E.g. brain tumors will clearly not have sufficient conspicuity. Despite the selection of the acquisition parameters, it is expected that the segmentation algorithms will be able to handle the breadth. |
| **2** | **Q. What kind of additional study (if any is needed) would best prove the profile claim?** A. Additional study (as described in the evolving Levels of Confidence document) would provide increased confidence. With this stabilized specification QIBA CT can proceed to such testing. |
| **3** | **Q. How do we balance specifying what to accomplish vs how to accomplish it?**E.g. if the requirement is that the scan be performed the same way, do we need to specify that the system or the Technologist record how each scan is performed? If we don’t, how will the requirement to “do it the same” be met?A: Have made revisions to text to try to achieve an appropriate balance. The details of compliance testing are still not complete and will require further work in future drafts of the profile. |
| **4** | **Q. Should there be a “patient appropriateness” or “subject selection” section?**A. The claim is conditioned upon the tumor being measurable (and criteria are listed) and a section describes characteristics of appropriate (and/or inappropriate) subjects.  |
| **5** | **Q. Does 4cm/sec “scan speed” preclude too many sites?** A. No.Most 16-slice (and greater) scanners would be able to achieve this (although due to an idiosyncracy of the available scan modes, the total collimation needs to be dropped to 16mm rather than 20mm)Some examples that would meet this include:(a) 16 x 1mm collimation with 0.5 second rotation time and pitch ³ 1.25 OR(b) 16 x 1mm collimation with 0.4 second rotation time and pitch ³ 1 OR (c) 16 x 1.25 mm collimation with 0.5 second rotation time and pitch ³ 1 OR(d) 16 x 1.5mm collimation with 0.5 second rotation time and pitch ³ .833Keep in mind that 16 x 0.75 mm collimation would require(i) pitch > 1.67 at 0.5 second rotation time (which breaks the Pitch< 1.5 requirement OR(ii) pitch > 1.33 at 0.4 second rotation time (which is fine)A 4cm/sec threshold is needed since it would likely alleviate potential breath hold issues. Because the reconstructed image thickness allowed here was > 2 mm, all of the above collimation settings would be able to meet both the breath hold requirements as well as the reconstructed image thickness requirements.  |
| **6** | **Q. What do we mean by noise and how do we measure it?**A. Noise means standard deviation of a region of interest as measured in a homogeneous water phantom.FDA has starting looking at Noise Power Spectrum in light of recent developments in iterative reconstruction and an interest in evaluating what that does to the image quality/characteristics. QIBA should follow what comes out of those discussions, but since FDA is not mandating it and since few systems or sites toda are in a position to measure or make effective use of it, this profile will not mandate it either. It has promise though and would be worth considering for future profile work. |
| **7** | **Q. Is 5HU StdDev a reasonable noise value for all organs?** A. No. Will change to 18HU.Not sure where the 5 HU standard deviation came from. The 1C project used a standard deviation of 18HU. At UCLA, our Siemens Sensation 64 will yield a standard deviation of 17 HU for: a. 120kVp, 50 eff. mAs, 1 mm thickness, B30F filterTo get this down to 5 HU would require: a. Increasing the eff. mAs to 550, OR b. Increasing the slice thickness to 2 mm AND increasing eff. mAs to 275 |
| **8** | **Q. Are there sufficient DICOM fields for all of what we need to record in the image header, and what are they specifically?** A. For those that exist, we need to name them explicitly. For those that may not currently exist, we need to work with the appropriate committees to have them added. |
| **9** | **Q. Have we worked out the details for how we establish compliance to these specifications?** A. See Section 4. |
| **10** | **Q. What is the basis of the specification of 15% for the variability in tumor volume assessment within the Image Analysis section, and is it inclusive or exclusive of reader performance?** A. For the basis, see the paragraph below the table in Section B.2. It includes reader performance.Allocation of variability across the pipeline (shown in Figure 1) is fraught with difficulty and accounting for reader performance is difficult in the presence of different levels of training and competence among readers. Input on these points to help with this is appreciated (as is also the case for all aspects of this Profile). |
| **11** | **Q. Should we specify all three levels (Acceptable, Target, Ideal) for each parameter?**A. No. As much as possible, provide just the Acceptable value. The Acceptable values should be selected such that the profile claim will be satisfied. |
| **12** | **Q. What is the basis for our claim, and is it only aspirational?**A. Our claim is informed by an extensive literature review of results achieved under a variety of conditions. From this perspective it may be said to be well founded; however, we acknowledge that the various studies have all used differing approaches and conditions that may be closer or farther from the specification outlined in this document. In fact the purpose of this document is to fill this community need. Until field tested, the claim may be said to be “consensus.” Commentary to this effect has been added in the Claims section, and the Background Information appendix has been augmented with the table summarizing our literature sources. |
| **13** | **Q. What about dose?**A. A discussion has been added in Section 2 to address dose issues.  |
| **14** | **Q. Are there any IRB questions that should be addressed?**A. The UPICT protocol that will be derived from this Profile will flush out any IRB issues if they exist. |
| **15** | **Q. What mechanisms are suggested to achieve consistency with baseline parameters?**A. Basically manual for now.In the future we can consider requiring the parameters be stored in the DICOM image headers or (future) DICOM Protocol Objects, and require systems be able to query/retrieve/import such objects to read prior parameters.  |
| **16** | **Q. Should the claim (and profile) reflect reproducibility (actors must be compliant but are allowed to be different) or repeatability (actors must be compliant and must be the same)?**A. State claim for scanner/reader/analysis-SW all permitted to be different across timepoints.This is most applicable to clinical practice. Although QIBA started by looking at Clinical Trials, it has really evolved to address Clinical Practice and that is more generally useful and practical.Different scanners cannot be avoided. Theoretically, different readers/SW could be avoided by requiring re-read/re-analyze of prior timepoints if different, but practically speaking, routine practice will not accommodate re-reading.Note that when actors are not different across timepoints you are still compliant with the profile and performance can be expected to improve. If we can provide informative material about the degree of improvement, that would be helpful for some users. If there is minimal additional load in terms of assessment procedures, we can also consider elevating such alternate scenario performance to be part of the claim too. |
| **17** | Should assessment procedures be "open book" or "closed book"?A: "Open book" for now.With “closed book” the correct answers are not available to the assessor. This depends on someone setting up infrastructure for manufacturers/sites to submit data and a system to calculate and return a “closed book” score. May consider this in the future if there is sufficient need/value. |

# Appendices

## Appendix A: Acknowledgements and Attributions

This document is proffered by the Radiological Society of North America (RSNA) Quantitative Imaging Biomarker Alliance (QIBA) Volumetric Computed Tomography (CTVol) Technical Committee. The CTVol technical committee is composed of scientists representing the imaging device manufacturers, image analysis software developers, image analysis laboratories, biopharmaceutical industry, academia, government research organizations, professional societies, and regulatory agencies, among others. All work is classified as pre-competitive.

A more detailed description of the committee and its work can be found at the following web link: http://qibawiki.rsna.org/index.php?title=Volumetric\_CT.

The Volumetric CT Technical Committee (in alphabetical order):

* Athelogou, M. Definiens AG
* Avila, R. Kitware, Inc.
* Beaumont, H. Median Technologies
* Borradaile, K. Core Lab Partners
* Buckler, A. BBMSC
* Clunie, D. Core Lab Partners
* Cole, P. Imagepace
* Conklin, J. ICON Medical Imaging
* Dorfman, GS. Weill Cornell Medical College
* Fenimore, C. Nat Inst Standards & Technology
* Ford, R. Princeton Radiology Associates.
* Garg, K. University of Colorado
* Garrett, P. Smith Consulting, LLC
* Goldmacher, G. ICON Medical Imaging
* Gottlieb, R. University of Arizona
* Gustafson, D. Intio, Inc.
* Hayes, W. Bristol Myers Squibb
* Hillman, B. Metrix, Inc.
* Judy, P. Brigham and Women’s Hospital
* Kim, HJ. University of California Los Angeles
* Kohl, G. Siemens AG
* Lehner, O. Definiens AG
* Lu, J. Nat Inst Standards & Technology
* McNitt-Gray, M. University California Los Angeles
* Mozley, PD. Merck & Co Inc.
* Mulshine, JL. Rush University
* Nicholson, D. Definiens AG
* O'Donnell, K. Toshiba Medical Research Institute - USA
* O'Neal, M. Core Lab Partners
* Petrick, N. US Food and Drug Administration
* Reeves, A. Cornell University
* Richard, S. Duke University
* Rong, Y. Perceptive Informatics, Inc.
* Schwartz, LH. Columbia University
* Saiprasad, G. University of Maryland
* Samei, E. Duke University
* Siegel, E. University of Maryland
* Silver, M. Toshiba Medical Research Institute – USA
* Steinmetz, N. Translational Sciences Corporation
* Sullivan, DC. RSNA Science Advisor and Duke University
* Tang, Y. CCS Associates
* Thorn, M. Siemens AG
* Vining, DJ. MD Anderson Cancer Center
* Yankellivitz, D. Mt. Sinai School of Medicine
* Yoshida, H. Harvard MGH
* Zhao, B. Columbia University

The Volumetric CT Technical Committee is deeply grateful for the support and technical assistance provided by the staff of the Radiological Society of North America.

## Appendix B: Background Information

## B.1 QIBA

The Quantitative Imaging Biomarker Alliance (QIBA) is an initiative to promote the use of standards to reduce variability and improve performance of quantitative imaging in medicine. QIBA provides a forum for volunteer committees of care providers, medical physicists, imaging innovators in the device and software industry, pharmaceutical companies, and other stakeholders in several clinical and operational domains to reach consensus on standards-based solutions to critical quantification issues. QIBA publishes the specifications they produce (called QIBA Profiles), first to gather public comment and then for field test by manufacturers and users.

QIBA envisions providing a process for developers to test their implementations of QIBA Profiles through a compliance mechanism. Purchasers can specify conformance with appropriate QIBA Profiles as a requirement in Requests For Proposals (RFPs). Manufacturers who have successfully implemented QIBA Profiles in their products can publish QIBA Conformance Statements. The Conformance Statements are accompanied by “Model-specific Parameters” (as shown in Appendix D) describing how to configure their product for alignment with the Profile.

General information about QIBA, including its governance structure, sponsorship, member organizations and work process, is available at <http://qibawiki.rsna.org/index.php?title=Main_Page>.

QIBA has constructed a systematic approach for standardizing and qualifying volumetry as a biomarker of response to treatments for a variety of medical conditions, including cancers in the lung (either primary cancers or cancers that metastasize to the lung [18]).

## B.2 CT Volumetry for Cancer Response Assessment: Overview and Summary

X-ray computed tomography provides an effective imaging technique for assessing treatment response in subjects with cancer. Size quantification is helpful to evaluate tumor changes over the course of illness. Currently most size measurements are uni-dimensional estimates of longest diameters (LDs) on axial slices, as specified by RECIST (Response Evaluation Criteria In Solid Tumors). Since its introduction, limitations of RECIST have been reported. Investigators have suggested that quantifying whole tumor volumes could solve some of the limitations of diameter measures [1-2] and many studies have explored the value of volumetry [3-12] . This document proposes standardized methods for performing repeatable volume measurements.

Anatomic imaging using computed tomography (CT) has been historically used to assess tumor burden and to determine tumor response to treatment (or progression) based on uni-dimensional or bi-dimensional measurements. The original WHO response criteria were based on bi-dimensional measurements of the tumor and defined response as a decrease of the sum of the product of the longest perpendicular diameters of measured tumors by at least 50%. The rationale for using a 50% threshold value for definition of response was based on data evaluating the reproducibility of measurements of tumor size by palpation and on planar chest x-rays [[33](#_ENREF_33), [34](#_ENREF_34)]. The more recent RECIST criteria introduced by the National Cancer Institute (NCI) and the European Organisation for Research and Treatment of Cancer (EORTC) standardized imaging techniques for anatomic response assessment by specifying minimum size thresholds for measurable tumors and considered other imaging modalities beyond CT. As well, the RECIST criteria replace longest bi-directional diameters with longest uni-dimensional diameter as the representation of a measured tumor [[35](#_ENREF_35)]. RECIST defines response as a 30% decrease of the largest diameter of the tumor. For a spherical tumor, this is equivalent to a 50% decrease of the product of two diameters. Current response criteria were designed to ensure a standardized classification of tumor shrinkage after completion of therapy. They have not been developed on the basis of clinical trials correlating tumor shrinkage with patient outcome.

Technological advances in signal processing and the engineering of multi-detector row computed tomography (MDCT) devices have resulted in the ability to acquire high-resolution images rapidly, resulting in volumetric scanning of anatomic regions in a single breath-hold. Volume measurements may be a more sensitive technique for detecting longitudinal changes in tumor masses than linear tumor diameters as defined by RECIST. Comparative analyses in the context of clinical trial data have found volume measurements to be more reliable, and often more sensitive to longitudinal changes in size and thus to treatment response, than the use of a uni-dimensional diameter in RECIST. As a result of this increased detection sensitivity and reliability, volume measurements may improve the predictability of clinical outcomes during therapy compared with RECIST. Volume measurements could also benefit patients who need alternative treatments when their disease stops responding to their current regimens [[36-39](#_ENREF_36)].

The rationale for volumetric approaches to assessing longitudinal changes in tumor burden is multi-factorial. First, most cancers may grow and regress irregularly in three dimensions. Measurements obtained in the transverse 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 believed to be less subject to either reader error or inter-scan variations. For example, partial response using the RECIST criteria requires a greater than 30% decrease in tumor diameter, which corresponds to greater than 50% decrease in tumor volume. If one assumes a 21 mm diameter spherical tumor (of 4.8 cc volume), partial response 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 1.7 cc. The much greater absolute magnitude of volumetric changes is potentially less prone to measurement error than changes in diameter, particularly if the tumors are spiculated or otherwise irregularly shaped. As a result of the observed increased sensitivity and reproducibility, volume measurements may be more suited than uni-dimensional measurements to identify early changes in patients undergoing treatment.

**Table B.1 Summarizing the precision/reproducibility of volumetric measurements from clinical studies reported in the literature**

| **Scan** | **Reader** | **# of Readers** | **# of Patients** | **# of Nodules** | **Tumor Size,** **Mean (range)** | **Organ System** | **Volumetry,****95% CI of Measurement Difference** | **Volumetry, Measurement Difference %** | **1D Measurement, 95% CI of Measurement Difference** | **1D, Mean Measurement Difference %** | **Slice Thickness /Recon Interval, mm** | **Author, Year** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| repeat scans  | intra-reader | 1 | 20 | 218 | 9.85 mm | lung, mets |  -21.2 to 23.8%  | 1.30% |  |  | 1.0/0.7 | Gietama *et al*. 2007 [[40](#_ENREF_40)] |
| repeat scans  | intra-reader | 3 | 32 | 32 | 38 mm (11–93 mm) | lung, NSCLC |  -12 to 13.4% | 0.70% |  -7.3% to 6.2% | -0.60% | 1.25/1.25 | Zhao *et al*. 2009 [[8](#_ENREF_8)] |
| same scan | intra-reader | 1 | 10 | 50 | 6.9 mm (2.2–20.5 mm) | lung, mets |  -3.9 to 5.7% | 0.90% | not reported | not reported | 1.25/0.8 | Wormanns *et al*. 2004 [[41](#_ENREF_41)] |
| same scan | inter-reader | 2 | 10 | 50 | 6.9 mm (2.2–20.5 mm) | lung, mets |  -5.5 to 6.6% | 0.50% | not reported | not reported | 1.25/0.8 | Wormanns *et al*. 2004 [[41](#_ENREF_41)] |
| repeat scans  | not specified | not specified | 10 | 151 | 7.4 (2.2–20.5 mm) | lung, mets |  -20.4 to 21.9% | 1.50% | not reported | not reported | 1.25/0.8 | Wormanns *et al*. 2004 [[41](#_ENREF_41)] |
| repeat scans  | not specified | not specified | 10 | 105 |  <10 mm | lung, mets |  -19.3 to 20.4% | 1.70% | not reported | not reported | 1.25/0.8 | Wormanns *et al*. 2004 [[41](#_ENREF_41)] |
| same scan (5 sets, 1 set/phase)  | intra-reader ? (consensus by 2 readers), 3 x reading | 2 | 30 | 73 | ~1–9 mm [25.3 (0.2–399 mm3)] | lung, noncalcified nodules | coefficient of variance as large as 34.5% (95% CI not reported) | not reported | not reported | not reported | 0.75/0.6 | Boll *et al*. 2004 [[42](#_ENREF_42)] |
| same scan  | inter-reader | 2 | 33 | 229 | 10.8 mm (2.8–43.6 mm), median 8.2 mm | lung, primary or mets |  -9.4 to 8.0% | 0.70% |  -31.0 to 27% | -2.00% | 1.0/0.8 | Hein *et al*. 2009 [[43](#_ENREF_43)] |
| same scan | inter-reader, inter-algorithms (6 readers x 3 algorithms) | 6 | 16 | 23 | not reported | lung, nodules |  55% (upper limit) | not reported | not reported | not reported | 1.25/0.625 | Meyer *et al*. 2006 [[44](#_ENREF_44)] |
| same scan | intra-reader | 2 | 50 | 202 | 3.16–5195 mm3, median 182.22 mm3 | lung, mets | % not reported | 0.15 to 0.22% | % not reported | 2.34–3.73% (p<0.05 1D vs 3D)  | 0.75/0.70 | Marten *et al*. 2006 [[45](#_ENREF_45)] |
| same scan | inter-reader | 2 | 50 | 202 | 3.16–5195 mm3, median 182.22 mm3 | lung, mets | % not reported | 0.22 to 0.29% | % not reported | 3.53–3.76% (p<0.05 1D vs 3D) | 0.75/0.70 | Marten *et al*. 2006 [[45](#_ENREF_45)] |
| same scan | inter-reader | 2 | 2239 | 4225 | 15–500 mm3 (effective diameter 3.1–9.8 mm) | lung, nodules |  -13.4 to 14.5% | 0.50% | not reported | not reported | 1.0/0.7 | Wang *et al*. 2008 [[1](#_ENREF_1)] |
| same scan | intra-reader | 2 | 24 | 52 | 8.5 mm (<5 to 18 mm) | lung, noncalcified nodules | 8.9 % (upper limit) | not reported | not reported | not reported | 1.25 or 2.5/not specified | Revel *et al*. [[46](#_ENREF_46)] |
| same scan | inter-reader (3 readers x 3 measurements) | 3 | 24 | 52 | 8.5 mm (< 18 mm) | lung, noncalcified nodules | 6.38 % (upper limit) | not reported | not reported | not reported | 1.25 or 2.5/not specified | Revel *et al*. [[46](#_ENREF_46)] |

**Abbreviations:** 1D = unidimensional; mets = metastasis; CI = confidence interval

The above table provides a basis for the 30% value in the Profile Claim. The range between the minimum and maximum values in the 95% CI of the measurement difference column is mostly within +/- 15%. Considering a large study from Wang et al using 2239 patients [[1](#_ENREF_1)], the 95% confidence interval ranged [-13.4%, 14.5%]. Thus, 30% is a conservative threshold of measurement variation. For example, the 30% change in the claim is the outside of 95% confidence interval of 15% of measurement variability when sample size is 40 or more.

## B.3 Detailed Literature Review by Indication

To date, volumetry has been evaluated in lung, liver, head and neck, esophagus, and rectal cancers, sarcoma, and lymphoma (Appendix 1, Tables 1–7). Most studies compared volumetry with either unidimensional RECIST or bidimensional WHO classifications. Volumetry showed a high degree of concordance with uni- or bidimensional assessment in some studies [[47](#_ENREF_47), [48](#_ENREF_48)]; others showed considerable discordance between these methods in response classifications [[49-52](#_ENREF_49)]. Correlation of volumetric assessment with pathologic response was examined in four studies (two esophageal, one gastric cancer, and one sarcoma) in patients who had or were having neoadjuvant chemotherapy. Among those four studies, volumetric assessment was correlated with pathologic response in two studies (one esophageal and one gastric study) [[53](#_ENREF_53), [54](#_ENREF_54)], whereas no such correlation was found in an esophageal study [[55](#_ENREF_55)] and a sarcoma study [[56](#_ENREF_56)]. Two of the above neoadjuvant studies also followed esophageal cancer patients for OS or PFS, but neither study showed correlation with volumetric assessment [[54](#_ENREF_54), [55](#_ENREF_55)]. In addition, two small studies [[57](#_ENREF_57), [58](#_ENREF_58)] with lymphoma patients showed that patients with greater reduction in tumor volume at 1–2 months of chemotherapy had lower probability of relapse at one year.

#### Lung Cancer (Tables B.2 and B.3)

Lung cancer typically spreads as it advances from localized disease to the neighboring lymph nodal structures of the lung (regional metastatic spread). The most advanced stage is metastasis to a distant site such as the brain or liver. In clinical trials, depending on the initial stage at diagnosis, either progression of localized disease or the discovery of a new site of metastatic dissemination is the basis for declaring failure of the efficacy of a new drug. In virtually all lung cancer clinical trials, there are situations when either a quantitative or a qualitative endpoint may be relevant, but it is likely that quantitative endpoints will be most frequently informative in trials.

With advanced disease, there is a tendency toward more frequent disease progression at a distant metastatic site rather than progression due to extension from the primary tumor [[59](#_ENREF_59)]. These patterns of disease progression impact clinical trial design in measuring drug response. However, there are exceptions to the pattern just described, such as bronchioalveolar carcinoma. This more indolent cancer tends to spread extensively within the lung but seldom to distant sites [[60](#_ENREF_60)].

**Staging, Therapeutic Options, and Response Assessment by Imaging Approaches**

Staging defines the extent of lung cancer dissemination at the time of initial patient diagnosis. The schema for staging lung cancer has been updated recently to more accurately cluster patients who benefit from particular therapeutic interventions with predictable outcomes [[61](#_ENREF_61)]. How staging relates to lung cancer drug therapy approaches, the imaging approaches used in those stages, and issues relative to image requirements is summarized in Table B.2 [[62](#_ENREF_62)].

**Table B.2. Summary of Image Processing Issues Relative to Lung Cancer Stage**

| **Stage** | **Percent of Cases** | **Percent 5-year Survival** | **Imaging Focus/ Therapy Focus** | **Imaging Tool** | **Issues** | **Thoracic Segmentation** | **Hi-Res** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| I | 16 | 49 | Primary tumor/ Neo and adjuvant RX | MDCT | Small cancers surrounded by air | Can be straightforward | Needed |
| II/III | 35 | 15.2 | Primary, hilar, and mediastinal lymph nodes/Combined modality | MDCT, PET | Larger tumors and nodes abut other structures | Often challenging | Optional |
| IV | 41 | 3 | Primary/regional nodes and metastatic sites/Chemotherapy | MDCT, PET, bone, brain scans | Tumor response often determined outside of the chest | Often challenging | Optional |

The imaging goal may vary in different disease stages. For example, with Stage IV lung cancer, the disease progression could be due to new growth in the primary lung tumor and/or metastasis of the cancer to a distant site, and not growth of the primary cancer site. In Stage II and III lung cancer, disease progression is often manifested by increased tumor involvement in regional lymph nodes. CT imaging would typically be used to assess potential disease progression in either the primary tumor or in the lymphatic tissue. The development of new sites of metastatic disease in a Stage IV clinical trial will require a different imaging approach. To assess for new sites of metastatic disease, CT may be used to look for thoracic, hepatic, or retroperitoneal sites of metastasis, and PET scans will frequently be used to assess the progression of metastatic disease across the entire body. Common both to improving size-based measures (*i.e.*, moving from linear diameters to volume) as well as more computationally sophisticated measures (*e.g.*, tissue density in CT, mechanistic measures in PET) is a need for means to qualify performance across stakeholders involved in the application of these measures.

The potential utility of volumetry in predicting treatment response in lung cancer patients has been investigated by several groups. Jaffe pointed out that the value of elegant image analysis has not been demonstrated yet in clinical trials [[12](#_ENREF_12)]. Value depends, at least in part, on the extent to which imaging endpoints meet criteria as substitute endpoints for clinical outcome measures. In this review, however, value is limited to the ability of imaging to predict either beneficial biological activity or progressive disease sooner than alternative methods of assessment, so that individual patients can move on to other treatment alternatives, or at the very least, stop being exposed to toxicity without benefit. In this context, value is predominantly a function of sensitivity and accuracy.

In 2006, Zhao and colleagues [[36](#_ENREF_36)] reported a study of 15 patients with lung cancer at a single center. They used MDCT scans with a slice thickness of 1.25 mm to automatically quantify unidimensional LDs, bidimensional cross products, and volumes before and after chemotherapy. They found that 11/15 (73%) of the patients had changes in volume of 20% or more, while only one (7%) and 4 (27%) of the sample had changes in uni- or bidimensional line-lengths of >20%. Seven (47%) patients had changes in volume of 30% or more; no patients had unidimensional line-length changes of 30% or more, and only two patients (13%) had changes in bidimensional cross products of 30% or more. The investigators concluded that volumetry was substantially more sensitive to drug responses than uni- or bidimensional line-lengths. However, this initial data set did not address the clinical value of increasing the sensitivity of change measurements.

In a follow-up analysis [[63](#_ENREF_63)], the same group used volumetric analysis to predict the biologic activity of epidermal growth factor receptor (EGFR) modulation in NSCLC, with EGFR mutation status as a reference. In this population of 48 patients, changes in tumor volume at three weeks after the start of treatment were found to be more sensitive and equally specific when compared to early diameter change at predicting EGFR mutation status. The positive predictive value of early volume response for EGFR mutation status in their patient population was 86%. The investigators concluded that early volume change has promise as an investigational method for detecting the biologic activity of systemic therapies in NSCLC.

In 2007, Schwartz and colleagues [[38](#_ENREF_38)] unidimensionally and volumetrically evaluated target lesions, including lymph node, liver, peritoneal, and lung metastases, in 25 patients with metastatic gastric cancer being treated with combination therapy, and reported that volumetry predicted clinical response earlier than unidimensional RECIST by an average of 50.3 days.

In 2008, Altorki and colleagues [[39](#_ENREF_39)] reported that volumetry is substantially more sensitive than changes in unidimensional diameters. In a sample of 35 patients with early-stage lung cancer treated with pazopanib, 30 of 35 (85.7%) were found to have a measurable decrease in tumor volume; only three of these 35 subjects met RECIST criteria for a PR.

In a retrospective analysis of 22 patients with locally advanced lung cancer treated with radiation and chemotherapy, assessment of treatment response by volume change was found to be in agreement with that by RECIST and WHO criteria (K 0.776; 95% CI 0.357–1.0 for agreement with both RECIST and WHO) [[48](#_ENREF_48)] in 21 of 22 patients.

In another retrospective analysis of 15 patients with lung metastases from colorectal cancer, renal cell, or breast carcinoma, volumetric assessment of 32 lung lesions at baseline and after 1–4 months standard chemotherapy or radiotherapy showed fair to poor agreement with either RECIST or WHO assessment for response classification [[49](#_ENREF_49)].

In another retrospective analysis of 68 patients with primary or metastatic lung malignancies, volumetric assessment of treatment response was found to be highly concordant with RECIST (K 0.79–0.87) and WHO assessment (K 0.83–0.84) [[47](#_ENREF_47)]. The intraobserver reproducibility of volumetric classification was 96%, slightly higher than that of RECIST and WHO. The relative measurement error of volumetric assessment was 8.97%, also slightly higher than that of unidimensional and bidimensional assessment.

In another retrospective analysis of nine patients with lung metastases who were undergoing chemotherapy, volumetric assessment of treatment response agreed in all but one case with RECIST assessment at the patient level (K 0.69); at the lesion level, volumetric and RECIST assessment agreed on 21 of the 24 lesions (K 0.75). The level of agreement between volumetric and RECIST assessment was equivalent or superior to that of inter-observer agreement using the RECIST criteria [[64](#_ENREF_64)].

#### Primary Liver Cancer and Metastatic Lesions in the Liver (Table B.4)

Hepatocellular carcinoma (HCC) is the most common form of liver cancer in adults [[65](#_ENREF_65)]. The majority of patients have underlying hepatic dysfunction, which complicates patient management and trial design in the search for effective treatment [[66](#_ENREF_66), [67](#_ENREF_67)]. Despite advances in many aspects of HCC treatment, >70% of HCC patients present with advanced disease and will not benefit from existing treatment modalities, including liver transplantation, surgical resection, and loco-regional therapies. At present, only one systemic agent, *i.e.*, sorafenib, is approved for advanced HCC patients. There remains a great need for safe and effective systemic therapies for HCC patients who progressed on or do not tolerate sorafenib and for patients with more advanced hepatic dysfunction. The liver is also a common site of metastatic spread; metastatic involvement of the liver can occur with many neoplasms, including lung, colorectal, esophageal, renal cell and breast, and stomach cancers, pancreatic carcinoma, and melanoma [[11](#_ENREF_11), [68](#_ENREF_68)].

Evidence that radiologic responses reflect clinical outcomes has recently emerged in patients who were receiving systemic therapy for advanced liver cancer. In a phase 3 trial, sorafenib, a small molecule kinase inhibitor, prolonged the survival of patients with advanced liver cancer to 10.7 months as compared with 7.9 months for the placebo group. The time to radiologic progression as defined by RECIST [[69](#_ENREF_69)] was also significantly prolonged in the sorafenib group, in parallel with the survival advantage [[70](#_ENREF_70)]. This survival advantage conferred by sorafenib was later confirmed in the Asian population [[71](#_ENREF_71)].

Volumetric CT has been investigated in only a few studies in patients with metastatic liver lesions [[51](#_ENREF_51), [72](#_ENREF_72)] or HCC [[73](#_ENREF_73)] (Appendix 1) as discussed below. These studies compared volumetry with RECIST and/or the bidimensional WHO method in classifying treatment response, and found considerable discordance between volumetry and RECIST or WHO assessment [[51](#_ENREF_51), [72](#_ENREF_72)].

Prasad and colleagues [[51](#_ENREF_51)] compared volumetric with unidimensional (RECIST) and bidimensional (WHO) measurements in assessing response to treatment in 38 patients with liver metastases from breast cancer in a phase 3 trial. PR was defined as >65% reduction in volume; PD was defined as >73% increase in volume; and stable disease was defined as changes in volume between those in PR and PD. Patients were treated with docetaxel or capecitabine plus docetaxel, and tumors were measured at baseline and six months posttreatment. Response assessment using uni- and bidimensional methods are highly concordant (37 of 38 patients). Volumetric assessment of tumor burden was discordant with uni- and bidimensional results in 12 (32%) and 13 (34%) patients, respectively.

In another retrospective analysis of 10 patients with liver metastases from colorectal (8), esophageal (1), and gastric (1) cancers who were receiving chemotherapy, 26 pairs of pre- and posttreatment CT scans were evaluated by bidimensional criteria (WHO) and volumetry. Stable disease in the volumetric analysis was defined as between an increase in volume of less than 40% and a reduction in volume of less than 65%. Discordance between the bidimensional assessment and volumetry was found in 19–35% of the cases in disease status categories [[72](#_ENREF_72)].

Stillwagon and colleagues [[73](#_ENREF_73)] used volumetric measurements to assess the response to radiation and chemotherapy in 194 patients with unresectable HCC. PD was defined as 25% increase in volume; PR was defined as 30% reduction in volume; and stable disease was defined as less than 25% increase or less than 30% decrease in tumor volume.

###

#### Lymphoma (Table B.5)

Lymphomas comprise ~30 distinct diseases. Volumetric assessment of lymphoma has been found to correlate with treatment outcome in two early studies [[57](#_ENREF_57), [58](#_ENREF_58)] using non-helical scanners. Agreement with RECIST and WHO assessment was also found to be excellent in another study [[74](#_ENREF_74)].

In a study of eight patients with Stage I and II diffuse large cell lymphoma of the mediastinum followed for 12 to 68 months (mean 29 months), tumor volume was assessed before and at 1 to 2 months after chemotherapy. The relative tumor volume reduction was higher in those who remained in remission than in patients who had relapsed (89% and 73% reduction, respectively). However, whether this difference was statistically significant was not reported. It was also noted that the initial tumor volume prior to chemotherapy was also greater in the group who later relapsed [[57](#_ENREF_57)].

In a study of 12 patients with stage IA to IIB mediastinal Hodgkin’s disease who were followed for 12 to 84 months (mean 35 months) after treatment, patients with a >85% reduction in volume at 1 to 2 months after six cycles of chemotherapy had a lower incidence of mediastinal relapse (0/6, 0%) compared with those having 85% of less reduction (4/6, 67%) [[58](#_ENREF_58)].

In a study of 16 patients with lymphoma or germ cell tumors, volumetric assessment of response to chemotherapy agreed completely with the WHO criteria in classifying responses of the lesions (20 lesions), and agreed in 18 of the 20 (90%) lesions with RECIST criteria [[74](#_ENREF_74)].

###

#### Colorectal and Gastric Cancers (Table B.6)

Data suggest that volumetry may be valuable in assessing response to neoadjuvant therapy in gastric and colorectal cancers. In a prospective phase 2 study in 33 patients with resectable advanced gastric cancer who had four cycles (eight weeks) of neoadjuvant chemotherapy before surgical resection, volume reduction of primary gastric cancer correlated with histopathologic grades of regression, but the unidimensional reduction of maximum thickness and standardized uptake value (SUV) of FDG-PET did not. The optimal cut-off value of the tumor volume reduction was determined to be 35.6%, resulting in a positive predictive value and negative predictive value of 69.9% and 100%, respectively [[53](#_ENREF_53)].

In a study of 15 patients with rectosigmoid cancer prospectively enrolled in neoadjuvant radiation therapy, using a reduction of >65% in tumor volume as the threshold for PR, volumetric analysis disagreed with the WHO criteria in classifying treatment response in one patient and with the RECIST assessment (measuring the maximal wall thickness) in four patients [[75](#_ENREF_75)].

###

#### Head and Neck Cancer (Table B.7)

Head and neck cancers are clinically heterogenous, comprising multiple anatomic sites of origin with distinct natural histories and prognoses. Cure rates are low (30–50%) in locally advanced disease.

The role of volumetry in response assessment in head and neck cancer is unclear. In two retrospective studies of 129 patients with early or late stages of oral cavity or oropharynx carcinoma, assessment of response by volumetry had low agreement (38–56%) with clinical assessment by inspection and palpation [[52](#_ENREF_52), [76](#_ENREF_76)]. In the first study of 42 patients with early-stage oral cavity or oropharynx carcinoma, volume assessment of response at three to four weeks after local chemotherapy had low agreement with clinical assessment by inspection and palpation according to WHO criteria (38%) in classifying treatment response. It is noted that the lesion volume was calculated manually, assuming lesions were ellipsoid-shaped [[52](#_ENREF_52)].

In the second retrospective study reported by the same group, 87 patients with advanced oral cavity or oropharynx carcinoma were assessed by lesion volume before and three weeks after local chemotherapy. Volume assessment of treatment response agreed with clinical assessment by WHO criteria in 49 of 87 patients (56%) [[76](#_ENREF_76)].

#### Sarcoma (Table B.8)

The response to treatment in sarcoma is difficult to objectively measure and quantify anatomically as shown by the limited usefulness of RECIST in this setting [[77](#_ENREF_77), [78](#_ENREF_78)]. Assessment of tumor dimensions in sites such as bone, bowel, and peritoneal metastases is problematic; in addition, tumor volume reductions that can be measured by standard criteria may occur slowly or not at all (*e.g.*, due to persistence of necrotic or fibrotic tissue).

Volumetry has not demonstrated a value in response assessment in sarcoma. In a study of 20 patients with locally advanced high-grade soft-tissue sarcoma prospectively enrolled in neoadjuvant therapy, volume assessment before and after pre-operative treatment failed to correlate with histopathologic response and was unable to differentiate histopathologic responders (n=6) from non-responders (n=14). In contrast, changes in FDG uptake measured by SUV (both mean and maximum) using PET were predictive of histopathologic response at a high accuracy (area under response operating characteristics (ROC) curve = 1.0 and 0.98, respectively) [[56](#_ENREF_56)].

**Table B.3.** **Evaluation of Response to Therapy by Volumetry in Lung Cancer**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Disease Stage/ Therapy**  | **Number of Patients Evaluated** | **VIA Response Measurement/Timing** | **Comparator** | **Results** | **Statistical Analysis** | **Reference** |
| NSCLC, locally advanced/radio ± chemo (mostly carboplatin/ paclitaxel | 22 | PR –65% | RECIST, WHO | Good concordance between 3D, 2D, 1D (cases). CR 4/4/4, PR 16/15/15, NR 2/3/3.  | Kappa values. 3D *vs* 2D Kappa 0.776(95% CI 0.357–1.0, substantial agreement); 3D *vs* 1D Kappa 0.776 (95% CI 0.357–1.0, substantial agreement); 1D *vs* 2D Kappa 1.0 (perfect agreement) | Werner-Wasik *et al*. 2001[[48](#_ENREF_48)] |
| NSCLC, early stage gefitinib 3 wks, neoadjuvant | 48 | –24.9% (dichotomizing cut-off) | EGFR mutation sensitizing tumor to tyrosine kinase inhibitor; volume change -65% (RECIST deduced); optimal cut-off 1D (–7%) | Optimal cut-off of 3D changes 24.9%; sensitivity 90%, specificity 89% for classifying tumor w/o EGFR sensitizing mutation; PPV 86%, NPV 92%. 3D (24.9%) superior to 1D (optimal and RECIST). | Youdens' index (sensitivity + specificity −1) for determination of optimal dichromatic cut-off value; Wilcoxon rank-sum test for significance of difference | Zhao *et al* 2010 [[63](#_ENREF_63)] |
| Lung mets from colorectum, renal cell, breast; standard chemo or radio | 15 | Stable disease -65% to +44%; 2 follow-ups, at 1–4 months | RECIST, WHO | Kappa 3D *vs* 1D 0.818 (Visit 1 to V2), 0.429 (V2 to V3); 3D *vs* 2D 0.412 (V1 to V2), 0.118 (V2 to V3); fair agreement 3D *vs* 1D; poor 2D *vs* 3D | Kappa values | Tran *et al* 2004 [[49](#_ENREF_49)] |
| NSCLC (16), SCLC (9), lung mets of various origins (43); treatment not specified | 68 | Stable disease –65% to +44%; 3 months for lung cancer, time varied for mets | RECIST, WHO | Kappa 1D *vs* 3D 0.79-0.87, Kappa 2D *vs* 3D 0.83-0.84 | Kappa values | Sohns *et al*. 2010 [[47](#_ENREF_47)] |
| Lung mets, unspecified origin; chemo | 9 (24 nodules) | Stable disease –65% to +73%;  | RECIST | At nodule/lesion level, disagreement 3 in 24 nodules (Kappa 0.75); at patient level, disagree 1/9 (Kappa 0.59) | Kappa values | Fraioli *et al*. 2006 [[64](#_ENREF_64)] |
| NSCLC, stage I or II, operable and resectable/ gefitinib > 21 days | 15 | –20% and –30%; 26.4 days since baseline scan | RECIST and WHO  | 3D more sensitive in detecting changes. > –20%: 3D: 11/15 (73%); 1D 1/15 (7%) (p< .01); 2D 4/15 (27%)(P= .04); > –30%: 3D, 7/15 (47%); 1D 0/15 (p=.02); 2D, 2/15 (13%) (p= .06). | P values | Zhao *et al*. 2006 [[36](#_ENREF_36)] |
| Mets to lymph node, liver, peritoneal and lung originated from primary gastric cancer or Gastroespoha-geal junction adenocarcinoma/irinotecan, cisplatin and bevacizumab | 25 | 3D, –65%/ 6-week follow-up for 10 cycles. 1D and 3D comparison made at the time with maximal clinical response | RECIST | 8/25 (72%) responders by both RECIST and 3D; 3D identified responders a mean of 50.3 days earlier than did RECIST criteria | There was a statistically significant (p<0.01) change in ratio of volume measurement change to RECIST measurement change for responding versus stable patients. | Schwartz *et al.* 2007 [[38](#_ENREF_38)] |
| NSCLC, Stage I/IIResectable/ neoadjuvant, pazopanib 800mg qd for 2 to 6 weeks | 35 | Volume change, response criteria not specified/1 week after last dose | RECIST | 3D: 30/36 (86%) –1% to –86%. 2/35 > –50% (86% and 75%; 23/35 (66%) > –10%; 12/35 > –30%; 1D 3/25 PR (reduction 86%, 75%, and 36%). Discordance between 3D and RECIST, not head-to-head comparison in % change. 3D superiority unclear. | Not specified | Altorki *et al.* 2010 [[39](#_ENREF_39)] |

 **Table B.4.** **Evaluation of Response to Therapy by Volumetry in Liver Cancer**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Disease Stage/ Therapy**  | **Number of Patients Evaluated** | **VIA Response Measurement/Timing** | **Comparator** | **Results** | **Statistical Analysis** | **Reference** |
| Hepatic mets from GIsystemic chemo +/- intra-arterial chemo | 10 (37 scans) | Stable disease –65% to +40% | WHO  | Discordance between 2D and 3D: 19% and 35% (w/o or w 5% variation interval). Conclusion: 2D and 3D not interchangable. 2D tended to identify PD when 3D indicated no change. All lesions hypodense. | Not specified | Garant *et al.* 1999 [[72](#_ENREF_72)] |
| Hepatic mets from breastdocetaxel *vs* capecitabine + docetaxel | 38 | Stable disease –65% to +73% | RECIST, WHO | Treatment response concordance 1D and 2D; discordance 1D *vs* 3D, and 2D *vs* 3D | Not specified | Prasad *et al.* 2000 [[51](#_ENREF_51)] |

 **Table B.5.** **Evaluation of Response to Therapy by Volumetry in Lymphoma**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Disease Stage/ Therapy**  | **Number of Patients Evaluated** | **VIA Response Measurement/Timing** | **Comparator** | **Results** | **Statistical Analysis** | **Reference** |
| Lymphoma or germ cell; chemo | 16 (20 lesions) | Volume change/timing not specified | RECIST, WHO | 3D agreed completely with 2D, agreed in 18/20 (lesions) with 1D | Coefficient of variation calculated | Sohaib *et al.* 2000 [[74](#_ENREF_74)] |
| Diffuse large cell lymphoma of the mediastinum; multiagent chemo | 8 | Volume change; 1–2 months (CT follow-up) | Relapse/ remission/ death | Patients were followed for minimum 1 yr or until death, mean 29 months (13–68 months). Reduction of tumor volume greater in pts in remission than in relapse (89% *vs* 73%, respectively).  | No statistical analysis performed | Willett *et al.* 1988 [[57](#_ENREF_57)] |
| Mediastinal Hodgkin's, stage IA to IIB; multiagent chemo | 12 | Volume change; 1–2 months (CT follow-up) | Relapse/ remission/ death | Patients were followed for minimum 1 yr or until death, mean 35 months (12–84 months). a >85% reduction in volume at 1 to 2 months after six cycles of chemotherapy had a lower incidence of mediastinal relapse (0/6, 0%) compared with those having 85% of less reduction (4/6, 67%)  | No statistical analysis performed | Willett *et al.* 1988 [[58](#_ENREF_58)] |

 **Table B.6.** **Evaluation of Response to Therapy by Volumetry in Colorectal and Gastric Cancers**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Disease Stage/ Therapy**  | **Number of Patients Evaluated** | **VIA Response Measurement/Timing** | **Comparator** | **Results** | **Statistical Analysis** | **Reference** |
| Gastric, resectable advanced; chemo, neoadjuvant: oxaliplatin, infusional 5-FU, leucovorin; 8 wks (4 cycles) | 33 | Volume change cut-off 35.6%; 4, 8 wks | Pathologic response | Volume reduction of primary tumor or index node correlated best with histopathologic grades for regression, followed by short diameter of index node. No correlation with thickness of primary gastric cancer, or SUV by PET/CT. The optimal cut-off value of the tumor volume reduction 35.6%.  | Spearman rank for correlation determination; ROC for determination of optimal cut-off value | Lee *et al.* 2009 [[53](#_ENREF_53)] |
| Rectosigmoid; neoadjuvant radiation | 15 | PR –65%; timing not specified | Maximal wall thickness (RECIST), WHO | Discordance w RECIST and WHO (4/15 and 1/15, respectively) | Student’st test for paired data; Pearson’s correlationtest. p < 0.05 | Luccichenti *et al.* 2005 [[75](#_ENREF_75)] |

 **Table B.7.** **Evaluation of Response to Therapy by Volumetry in Head and Neck Cancer**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Disease Stage/ Therapy**  | **Number of Patients Evaluated** | **VIA Response Measurement/Timing** | **Comparator** | **Results** | **Statistical Analysis** | **Reference** |
| Oral cavity and oropharynx, carcinoma T1/2; chemo (cisplatin), intra-arterial | 42 | CR –90%, PR -50%, stable disease –50% to +25%, PR >+25%; 4 wks | Clinical inspection and palpation of lesions, classified per WHO criteria | Concordance in classifying response categories 16 of 42 pts (38%) | Not reported for concordance analysis | Rohde 2006 [[52](#_ENREF_52)] |
| Oral cavity and oropharynx, carcinoma T3/4; chemo (cisplatin), intra-arterial | 87 | CR –90%, PR –50%, stable disease –50% to +25%, PR >+25%; 4 wks | Clinical inspection and palpation of lesions, classified per WHO criteria | Concordance in classifying response categories 49/87 pts (56%); Kappa value was not reported.  | Kappa for agreementbetween clinical and radiological remission rates | Rohde 2007 [[76](#_ENREF_76)] |

 **Table B.8.** **Evaluation of Response to Therapy by Volumetry in Sarcoma**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Disease Stage/ Therapy** | **Number of Patients Evaluated** | **VIA Response Measurement/Timing** | **Comparator** | **Results** | **Statistical Analysis** | **Reference** |
| Sarcoma, locally advanced high-grade, soft tissue; chemo (ifosfamide/doxorubicin or gemcitabine/docetaxel)± radiation | 20 | Volume change/timing not specified | Pathologic response | Volume reduction not significant pre- and post-treatment; not predictive of histopathologic response (6 responders, 14 nonresponders), AUC = 0.48 | ROC curve | Benz 2008 [[56](#_ENREF_56)] |

**Abbreviations**:

1D = unidimensional measurement; 2D = bidimensional measurement; 3D = volumetric measurement; AUC = area under the curve; CI = confidence interval; CR = complete response; EGFR = epidermal growth factor receptor; FU = fluorouracil; Mets = metastasis; NSCLC = non small cell lung cancer; OS = overall survival; PFS = progression free survival; PR = partial response; PR = partial response; RECIST = Response Evaluation Criteria in Solid Tumors; ROC = response operating characteristics; SCLC = small cell lung cancer.

## B.4 References

1. Wang, Y., R.J. van Klaveren, H.J. van der Zaag-Loonen, G.H. de Bock, 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): p. 625-31.

2. Buckler, A.J., P.D. Mozley, L. Schwartz, N. Petrick, et al., *Volumetric CT in lung cancer: an example for the qualification of imaging as a biomarker.* Academic radiology, 2010. 17(1): p. 107-15.

3. Mozley, P.D., L.H. Schwartz, C. Bendtsen, B. Zhao, N. Petrick, and A.J. Buckler, *Change in lung tumor volume as a biomarker of treatment response: a critical review of the evidence.* Ann Oncol, 2010. 21(9): p. 1751-5.

4. Buckler, A.J., *A procedural template for the qualification of imaging as a biomarker, using volumetric CT as an example*, in *IEEE Applied Imagery Pattern Recognition Workshop*2009: Cosmos Club, Washington, D.C. p. 7.

5. Buckler, A.J., J.L. Mulshine, R. Gottlieb, B. Zhao, P.D. Mozley, and L. Schwartz, *The use of volumetric CT as an imaging biomarker in lung cancer.* Academic radiology, 2010. 17(1): p. 100-6.

6. Buckler AJ, S.L., Petrick N, McNitt-Gray M, Zhao B, Fenimore C, Reeves AP, Mozley PD, Avila RS, *Data Sets for the Qualification of CT as a Quantitative Imaging Biomarker in Lung Cancer.* Optics express, 2010. 18(14): p. 16.

7. Shankar, L.K., A. Van den Abbeele, J. Yap, R. Benjamin, S. Scheutze, and T.J. Fitzgerald, *Considerations for the use of imaging tools for phase II treatment trials in oncology.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2009. 15(6): p. 1891-7.

8. Zhao, B., L.H. Schwartz, and S.M. Larson, *Imaging surrogates of tumor response to therapy: anatomic and functional biomarkers.* J Nucl Med, 2009. 50(2): p. 239-49.

9. Maitland, M.L., *Volumes to learn: advancing therapeutics with innovative computed tomography image data analysis.* Clin Cancer Res, 2010. 16(18): p. 4493-5.

10. Nishino, M., D.M. Jackman, H. Hatabu, P.A. Janne, B.E. Johnson, and A.D. Van den Abbeele, *Imaging of lung cancer in the era of molecular medicine.* Acad Radiol, 2011. 18(4): p. 424-36.

11. Koshariya, M., R.B. Jagad, J. Kawamoto, P. Papastratis, H. Kefalourous, T. Porfiris, C. Tzouma, and N.J. Lygidakis, *An update and our experience with metastatic liver disease.* Hepatogastroenterology, 2007. 54(80): p. 2232-9.

12. Jaffe, C.C., *Measures of response: RECIST, WHO, and new alternatives.* J Clin Oncol, 2006. 24(20): p. 3245-51.

13. Gavrielides, M.A., L.M. Kinnard, K.J. Myers, and N. Petrick, *Noncalcified lung nodules: volumetric assessment with thoracic CT.* Radiology, 2009. 251(1): p. 26-37.

14. Petrick, N., H.J.G. Kim, D. Clunie, K. Borradaile, et al., *Evaluation of 1D, 2D and 3D nodule size estimation by radiologists for spherical and non-spherical nodules through CT thoracic phantom imaging*, in *SPIE 2011*.

15. Kinnard, L.M., M.A. Gavrielides, K.J. Myers, R. Zeng, J. Peregoy, W. Pritchard, J.W. Karanian, and N. Petrick, *Volume error analysis for lung nodules attached to pulmonary vessels in an anthropomorphic thoracic phantom.* Proc SPIE, 2008. 6915: p. 69152Q; doi:10.1117/12.773039.

16. Gavrielides, M.A., R. Zeng, L.M. Kinnard, K.J. Myers, and N. Petrick, *A template-based approach for the analysis of lung nodules in a volumetric CT phantom study.* Proc SPIE, 2009. 7260: p. 726009; doi:10.1117/12.813560.

17. Winer-Muram, H.T., S.G. Jennings, C.A. Meyer, Y. Liang, A.M. Aisen, R.D. Tarver, and R.C. McGarry, *Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations.* Radiology, 2003. 229(1): p. 184-94.

18. Ravenel, J.G., W.M. Leue, P.J. Nietert, J.V. Miller, K.K. Taylor, and G.A. Silvestri, *Pulmonary nodule volume: effects of reconstruction parameters on automated measurements--a phantom study.* Radiology, 2008. 247(2): p. 400-8.

19. Borradaile, K. and R. Ford, *Discordance between BICR readers.* Appl Clin Trials, 2010. Nov 1: p. Epub.

20. Gavrielides, M.A., R. Zeng, K.J. Myers, B. Sahiner, and N. Petrick, *Benefit of Overlapping Reconstruction for Improving the Quantitative Assessment of CT Lung Nodule Volume.* Acad Radiol, 2012.

21. Gavrielides, M.A., R. Zeng, L.M. Kinnard, K.J. Myers, and N. Petrick, *Information-theoretic approach for analyzing bias and variance in lung nodule size estimation with CT: a phantom study.* IEEE Trans Med Imaging, 2010. 29(10): p. 1795-807.

22. Gavrielides, M.A., L.M. Kinnard, K.J. Myers, J. Peregoy, W.F. Pritchard, R. Zeng, J. Esparza, J. Karanian, and N. Petrick, *A resource for the assessment of lung nodule size estimation methods: database of thoracic CT scans of an anthropomorphic phantom.* Opt Express, 2010. 18(14): p. 15244-55.

23. Das, M., J. Ley-Zaporozhan, H.A. Gietema, A. Czech, et al., *Accuracy of automated volumetry of pulmonary nodules across different multislice CT scanners.* Eur Radiol, 2007. 17(8): p. 1979-84.

24. Bolte, H., C. Riedel, S. Muller-Hulsbeck, S. Freitag-Wolf, G. Kohl, T. Drews, M. Heller, and J. Biederer, *Precision of computer-aided volumetry of artificial small solid pulmonary nodules in ex vivo porcine lungs.* Br J Radiol, 2007. 80(954): p. 414-21.

25. Cagnon, C.H., D.D. Cody, M.F. McNitt-Gray, J.A. Seibert, P.F. Judy, and D.R. Aberle, *Description and implementation of a quality control program in an imaging-based clinical trial.* Acad Radiol, 2006. 13(11): p. 1431-41.

26. Goodsitt, M.M., H.P. Chan, T.W. Way, S.C. Larson, E.G. Christodoulou, and J. Kim, *Accuracy of the CT numbers of simulated lung nodules imaged with multi-detector CT scanners.* Med Phys, 2006. 33(8): p. 3006-17.

27. Oda, S., K. Awai, K. Murao, A. Ozawa, Y. Yanaga, K. Kawanaka, and Y. Yamashita, *Computer-aided volumetry of pulmonary nodules exhibiting ground-glass opacity at MDCT.* AJR Am J Roentgenol, 2010. 194(2): p. 398-406.

28. McNitt-Gray, M.F., L.M. Bidaut, S.G. Armato, C.R. Meyer, et al., *Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error.* Transl Oncol, 2009. 2(4): p. 216-22.

29. Keil, S., C. Plumhans, F.F. Behrendt, S. Stanzel, M. Suehling, G. Muhlenbruch, A.H. Mahnken, R.W. Gunther, and M. Das, *Semi-automated quantification of hepatic lesions in a phantom.* Invest Radiol, 2009. 44(2): p. 82-8.

30. *Guidance for Industry: Standards for Clinical Trial Imaging Endpoints*, 2011, U.S. Department of Health and Human Services, Food and Drug Administration.

31. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1986. 1(8476): p. 307-10.

32. Lin, L.I., *A concordance correlation coefficient to evaluate reproducibility.* Biometrics, 1989. 45(1): p. 255-68.

33. Moertel, C.G. and J.A. Hanley, *The effect of measuring error on the results of therapeutic trials in advanced cancer.* Cancer, 1976. 38(1): p. 388-94.

34. Lavin, P.T. and G. Flowerdew, *Studies in variation associated with the measurement of solid tumors.* Cancer, 1980. 46(5): p. 1286-90.

35. Eisenhauer, E.A., P. Therasse, J. Bogaerts, L.H. Schwartz, et al., *New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1).* Eur J Cancer, 2009. 45(2): p. 228-47.

36. Zhao, B., L.H. Schwartz, C.S. Moskowitz, M.S. Ginsberg, N.A. Rizvi, and M.G. Kris, *Lung cancer: computerized quantification of tumor response--initial results.* Radiology, 2006. 241(3): p. 892-8.

37. Zhao, B., G.R. Oxnard, C.S. Moskowitz, M.G. Kris, et al., *A pilot study of volume measurement as a method of tumor response evaluation to aid biomarker development.* Clin Cancer Res, 2010. 16(18): p. 4647-53.

38. Schwartz, L.H., S. Curran, R. Trocola, J. Randazzo, D. Ilson, D. Kelsen, and M. Shah, *Volumetric 3D CT analysis - an early predictor of response to therapy.* J Clin Oncol, 2007. 25(18s): p. abstr 4576.

39. Altorki, N., M.E. Lane, T. Bauer, P.C. Lee, et al., *Phase II proof-of-concept study of pazopanib monotherapy in treatment-naive patients with stage I/II resectable non-small-cell lung cancer.* J Clin Oncol, 2010. 28(19): p. 3131-7.

40. Gietema, H.A., C.M. Schaefer-Prokop, W.P. Mali, G. Groenewegen, and M. Prokop, *Pulmonary nodules: Interscan variability of semiautomated volume measurements with multisection CT-- influence of inspiration level, nodule size, and segmentation performance.* Radiology, 2007. 245(3): p. 888-94.

41. Wormanns, D., G. Kohl, E. Klotz, A. Marheine, F. Beyer, W. Heindel, and S. Diederich, *Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility.* Eur Radiol, 2004. 14(1): p. 86-92.

42. Boll, D.T., R.C. Gilkeson, T.R. Fleiter, K.A. Blackham, J.L. Duerk, and J.S. Lewin, *Volumetric assessment of pulmonary nodules with ECG-gated MDCT.* AJR Am J Roentgenol, 2004. 183(5): p. 1217-23.

43. Hein, P.A., V.C. Romano, P. Rogalla, C. Klessen, A. Lembcke, V. Dicken, L. Bornemann, and H.C. Bauknecht, *Linear and volume measurements of pulmonary nodules at different CT dose levels - intrascan and interscan analysis.* Rofo, 2009. 181(1): p. 24-31.

44. Meyer, C.R., T.D. Johnson, G. McLennan, D.R. Aberle, et al., *Evaluation of lung MDCT nodule annotation across radiologists and methods.* Acad Radiol, 2006. 13(10): p. 1254-65.

45. Marten, K., F. Auer, S. Schmidt, G. Kohl, E.J. Rummeny, and C. Engelke, *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): p. 781-90.

46. Revel, M.P., C. Lefort, A. Bissery, M. Bienvenu, L. Aycard, G. Chatellier, and G. Frija, *Pulmonary nodules: preliminary experience with three-dimensional evaluation.* Radiology, 2004. 231(2): p. 459-66.

47. Sohns, C., J. Mangelsdorf, S. Sossalla, F. Konietschke, and S. Obenauer, *Measurement of response of pulmonal tumors in 64-slice MDCT.* Acta Radiol, 2010. 51(5): p. 512-21.

48. Werner-Wasik, M., Y. Xiao, E. Pequignot, W.J. Curran, and W. Hauck, *Assessment of lung cancer response after nonoperative therapy: tumor diameter, bidimensional product, and volume. A serial CT scan-based study.* Int J Radiat Oncol Biol Phys, 2001. 51(1): p. 56-61.

49. Tran, L.N., M.S. Brown, J.G. Goldin, X. Yan, R.C. Pais, M.F. McNitt-Gray, D. Gjertson, S.R. Rogers, and D.R. Aberle, *Comparison of treatment response classifications between unidimensional, bidimensional, and volumetric measurements of metastatic lung lesions on chest computed tomography.* Acad Radiol, 2004. 11(12): p. 1355-60.

50. Jennings, P., S. Aydin, J. Bennett, R. McBride, et al., *Inter-laboratory comparison of human renal proximal tubule (HK-2) transcriptome alterations due to Cyclosporine A exposure and medium exhaustion.* Toxicol In Vitro, 2009. 23(3): p. 486-99.

51. Prasad, S.R., K.S. Jhaveri, S. Saini, P.F. Hahn, E.F. Halpern, and J.E. Sumner, *CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques initial observations.* Radiology, 2002. 225(2): p. 416-9.

52. Rohde, S., A.F. Kovacs, J. Berkefeld, and B. Turowski, *Reliability of CT-based tumor volumetry after intraarterial chemotherapy in patients with small carcinoma of the oral cavity and the oropharynx.* Neuroradiology, 2006. 48(6): p. 415-21.

53. Lee, S.M., S.H. Kim, J.M. Lee, S.A. Im, et al., *Usefulness of CT volumetry for primary gastric lesions in predicting pathologic response to neoadjuvant chemotherapy in advanced gastric cancer.* Abdom Imaging, 2009. 34(4): p. 430-40.

54. Beer, A.J., H.A. Wieder, F. Lordick, K. Ott, M. Fischer, K. Becker, J. Stollfuss, and E.J. Rummeny, *Adenocarcinomas of esophagogastric junction: multi-detector row CT to evaluate early response to neoadjuvant chemotherapy.* Radiology, 2006. 239(2): p. 472-80.

55. Griffith, J.F., A.C. Chan, L.T. Chow, S.F. Leung, Y.H. Lam, E.Y. Liang, S.C. Chung, and C. Metreweli, *Assessing chemotherapy response of squamous cell oesophageal carcinoma with spiral CT.* Br J Radiol, 1999. 72(859): p. 678-84.

56. Benz, M.R., M.S. Allen-Auerbach, F.C. Eilber, H.J. Chen, S. Dry, M.E. Phelps, J. Czernin, and W.A. Weber, *Combined assessment of metabolic and volumetric changes for assessment of tumor response in patients with soft-tissue sarcomas.* J Nucl Med, 2008. 49(10): p. 1579-84.

57. Willett, C.G., M.A. Stracher, R.M. Linggood, L.M. Miketic, J.C. Leong, S.J. Skates, D.C. Kushner, and J.O. Jacobson, *Three-dimensional volumetric assessment of response to treatment: stage I and II diffuse large cell lymphoma of the mediastinum.* Radiother Oncol, 1988. 12(3): p. 193-8.

58. Willett, C.G., R.M. Linggood, J.C. Leong, L.M. Miketic, M.A. Stracher, S.J. Skates, and D.C. Kushner, *Stage IA to IIB mediastinal Hodgkin's disease: three-dimensional volumetric assessment of response to treatment.* J Clin Oncol, 1988. 6(5): p. 819-24.

59. Wakelee, H.A., P. Bernardo, D.H. Johnson, and J.H. Schiller, *Changes in the natural history of nonsmall cell lung cancer (NSCLC)--comparison of outcomes and characteristics in patients with advanced NSCLC entered in Eastern Cooperative Oncology Group trials before and after 1990.* Cancer, 2006. 106(10): p. 2208-17.

60. Gandara, D.R., D. Aberle, D. Lau, J. Jett, T. Akhurst, R. Heelan, J. Mulshine, C. Berg, and E.F. Patz, Jr., *Radiographic imaging of bronchioloalveolar carcinoma: screening, patterns of presentation and response assessment.* J Thorac Oncol, 2006. 1(9 Suppl): p. S20-6.

61. Goldstraw, P., J. Crowley, K. Chansky, D.J. Giroux, P.A. Groome, R. Rami-Porta, P.E. Postmus, V. Rusch, and L. Sobin, *The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours.* J Thorac Oncol, 2007. 2(8): p. 706-14.

62. Jemal, A., R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M.J. Thun, *Cancer statistics, 2008.* CA Cancer J Clin, 2008. 58(2): p. 71-96.

63. Zhao, B., G.R. Oxnard, C.S. Moskowitz, M.G. Kris, et al., *A Pilot Study of Volume Measurement as a Method of Tumor Response Evaluation to Aid Biomarker Development.* Clin Cancer Res, 2010. 16: p. 4647-4653.

64. Fraioli, F., L. Bertoletti, A. Napoli, F.A. Calabrese, R. Masciangelo, E. Cortesi, C. Catalano, and R. Passariello, *Volumetric evaluation of therapy response in patients with lung metastases. Preliminary results with a computer system (CAD) and comparison with unidimensional measurements.* Radiol Med, 2006. 111(3): p. 365-75.

65. Parkin, D.M., F. Bray, J. Ferlay, and P. Pisani, *Global cancer statistics, 2002.* CA Cancer J Clin, 2005. 55(2): p. 74-108.

66. Llovet, J.M. and J. Bruix, *Systematic review of randomized trials for unresectable hepatocellular carcinoma: Chemoembolization improves survival.* Hepatology, 2003. 37(2): p. 429-42.

67. Lopez, P.M., A. Villanueva, and J.M. Llovet, *Systematic review: evidence-based management of hepatocellular carcinoma--an updated analysis of randomized controlled trials.* Aliment Pharmacol Ther, 2006. 23(11): p. 1535-47.

68. Keil, S., F.F. Behrendt, S. Stanzel, M. Suhling, et al., *Semi-automated measurement of hyperdense, hypodense and heterogeneous hepatic metastasis on standard MDCT slices. Comparison of semi-automated and manual measurement of RECIST and WHO criteria.* Eur Radiol, 2008. 18(11): p. 2456-65.

69. Therasse, P., S.G. Arbuck, E.A. Eisenhauer, J. Wanders, et al., *New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada.* J Natl Cancer Inst, 2000. 92(3): p. 205-16.

70. Llovet, J.M., S. Ricci, V. Mazzaferro, P. Hilgard, et al., *Sorafenib in advanced hepatocellular carcinoma.* N Engl J Med, 2008. 359(4): p. 378-90.

71. Cheng, A.L., Y.K. Kang, Z. Chen, C.J. Tsao, et al., *Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial.* Lancet Oncol, 2009. 10(1): p. 25-34.

72. Garant, M., M. Trudeau, C. Reinhold, and P.M. Bret, *Liver metastasis: comparison of 2 methods for reporting of disease in patients receiving chemotherapy.* Can Assoc Radiol J, 1999. 50(1): p. 13-6.

73. Stillwagon, G.B., S.E. Order, C. Guse, J.L. Klein, P.K. Leichner, S.A. Leibel, and E.K. Fishman, *194 hepatocellular cancers treated by radiation and chemotherapy combinations: toxicity and response: a Radiation Therapy Oncology Group Study.* Int J Radiat Oncol Biol Phys, 1989. 17(6): p. 1223-9.

74. Sohaib, S.A., B. Turner, J.A. Hanson, M. Farquharson, R.T. Oliver, and R.H. Reznek, *CT assessment of tumour response to treatment: comparison of linear, cross-sectional and volumetric measures of tumour size.* Br J Radiol, 2000. 73(875): p. 1178-84.

75. Luccichenti, G., F. Cademartiri, M. Sianesi, L. Roncoroni, P. Pavone, and G.P. Krestin, *Radiologic assessment of rectosigmoid cancer before and after neoadjuvant radiation therapy: comparison between quantitation techniques.* AJR Am J Roentgenol, 2005. 184(2): p. 526-30.

76. Rohde, S., B. Turowski, J. Berkefeld, and A.F. Kovacs, *CT-based evaluation of tumor volume after intra-arterial chemotherapy of locally advanced carcinoma of the oral cavity: comparison with clinical remission rates.* Cardiovasc Intervent Radiol, 2007. 30(1): p. 85-91.

77. Therasse, P., *Measuring the clinical response. What does it mean?* Eur J Cancer, 2002. 38(14): p. 1817-23.

78. McHugh, K. and S. Kao, *Response evaluation criteria in solid tumours (RECIST): problems and need for modifications in paediatric oncology?* Br J Radiol, 2003. 76(907): p. 433-6.

#

## Appendix C: Conventions and Definitions

Acquisition vs. Analysis vs. Interpretation: This document organizes acquisition, reconstruction, post-processing, analysis and interpretation as steps in a pipeline that transforms data to information to knowledge. Acquisition, reconstruction and post-processing are considered to address the collection and structuring of new data from the subject. Analysis is primarily considered to be computational steps that transform the data into information, extracting important values. Interpretation is primarily considered to be judgment that transforms the information into knowledge. (The transformation of knowledge into wisdom is beyond the scope of this document.)

Image Analysis, Image Review, and/or Read: Procedures and processes that culminate in the generation of imaging outcome measures, such tumor response criteria. Reviews can be performed for eligibility, safety or efficacy. The review paradigm may be context specific and dependent on the specific aims of a trial, the imaging technologies in play, and the stage of drug development, among other parameters.

Image Header: that part of the image file (or dataset containing the image) other than the pixel data itself.

Imaging Phantoms: devices used for periodic testing and standardization of image acquisition. This testing must be site specific and equipment specific and conducted prior to the beginning of a trial (baseline), periodically during the trial and at the end of the trial.

Time Point: a discrete period during the course of a clinical trial when groups of imaging exams or clinical exams are scheduled.

Tumor Definition Variability: the clarity of the tumor boundary in the images. It originates from the biological characteristics of the tumor, technical characteristics of the imaging process, and perhaps on the perception, expertise and education of the operator.

Technical Variability - originates only from the ability to drawing unequivocal objects. In other words, the perception of tumor definition is supposed absolutely clear and similar for any given operator when attempting to assess “Technical” variability.

Global Variability - partitioned as the variability in the tumor definition plus the “Technical” variability.

Intra-Rater Variability - is the variability in the interpretation of a set of images by the same reader after an adequate period of time inserted to reduce recall bias.

Inter-Rater Variability - is the variability in the interpretation of a set of images by the different readers.

Repeatability – considers multiple measurements taken under the same conditions (same equipment, parameters, reader, algorithm, etc) but different subjects.

Reproducability – considers multiple measurements taken where one or more conditions have changed.

## Appendix D: Model-specific Instructions and Parameters

For acquisition modalities, reconstruction software and software analysis tools, Profile compliance requires meeting the Activity specifications above; e.g. in Sections 3.2, 3.3 and 3.5.

This Appendix provides, as an informative annex to the Profile, some specific acquisition parameters, reconstruction parameters and analysis software parameters that are expected to be compatible with meeting the Profile requirements. Just using these parameters without meeting the requirements specified in the Profile is not sufficient to achieve compliance. Conversely, it is possible to use different compatible parameters and still achieve compliance.

Additional parameter sets may be found in QIBA Conformance Statements published by manufacturers and sites. Manufacturers claiming product compliance with this QIBA Profile are required to provide such instructions and parameters describing the conditions under which their product achieved compliance.

Sites using models listed here are encouraged to consider these parameters for both simplicity and consistency. Sites using models not listed here may be able to devise their own settings that result in data meeting the requirements. Tables like the following may be used by sites that wish to publish their successful/best practices.

In any case, sites are responsible for adjusting the parameters as appropriate for individual subjects.

Discussion:

It would likely be useful to include a description of the imaging subject in the following tables.

In terms of standardization, it may make sense to ask manufacturers to publish parameters for a known reference phantom as a stable benchmark for sites to adjust for individual patient variations.

**Table D.1 Model-specific Parameters for Acquisition Devices**

| Acquisition Device | Settings Compatible with Compliance |
| --- | --- |
| <Manufacturer><Model><Version> | Submitted by: |
|

|  |  |
| --- | --- |
| kVp |  |
| Number of Data Channels (N) |  |
| Width of Each Data Channel (T, in mm) |  |
| Gantry Rotation Time in seconds |  |
| mA |  |
| Pitch |  |
| Scan FoV |  |

 |

**Table D.2 Model-specific Parameters for Reconstruction Software**

| Reconstruction Software | Settings Compatible with Compliance |
| --- | --- |
| <Manufacturer><Model><Version> | Submitted by: |
|

|  |  |
| --- | --- |
| Reconstructed Slice Width, mm |  |
| Reconstruction Interval |  |
| Display FOV, mm |  |
| Recon kernel |  |

 |

**Table D.3 Model-specific Parameters for Image Analysis Software**

| Image Analysis Software | Settings Compatible with Compliance |
| --- | --- |
| <Manufacturer><Model><Version> | Submitted by: |
|

|  |  |
| --- | --- |
| a |  |
| b |  |
| c |  |
| d |  |

 |

4.1.1 Parameter Derivation Procedure

This procedure can be used by a manufacturer or an imaging site to select an appropriate reconstruction kernel??? for an acquisition device model: and various parameter sets that might also be compliant (but it’s not the only way that’s acceptable to QIBA).

* Set the scanning field of view for the patient, the Kyoto Kagaku chest phantom. This setting is to be used to image the ACR phantom. *Special handling*: In scanning the ACR CT phantom, some manufacturers specify the use of a FOV appropriate to the ACR device (See: <http://www.acr.org/accreditation/computed/ct_faq.aspx#thirteen>). In this case, follow the manufacturer’s guidance for the ACR phantom. As an example, for the Aquilion16 scanner see the guidance in slide 11 of

<http://www.tams-media.com/tams2008/sales/faqs/CT_pdfs/ACR_Guide_Aquilion16.pdf>

* Set the beam voltage to 120 kVp
* Set the slice thickness to between 0.75 and 1.25 mm (depending on the available reconstructed slice thicknesses of the scanner)
* Set nominal beam collimation (NxT such as 16 x 0.5mm, or 128 x 0.6mm, 320 x 0.5 mm) rotation time and pitch such that scan can cover a 35 cm thorax in 15 seconds or less
* ITERATE (hopefully only a few times) on reconstruction kernels to meet spatial resolution spec.



Figure : Establishing spatial resolution

* ITERATE (again, hopefully just a few times) on mAs or effective mAs setting, given beam collimation, pitch and rotation time.



Figure : Establishing noise spec

* If the scanning FOV is to be changed for the scan of the lung phantom, reset the FOV accordingly and rescan the ACR phantom. Measure the quality parameters, the noise and resolution, with the changed settings.

The quality parameters are expected to change under a changed scanning field of view, as in the special handling. If this is the case, the comparability of the quality of the various scanners is lost. In the analysis of the quantitative measurements of nodule sizes, the data on the actual quality measures may prove to be useful in analyzing device differences.

## Appendix E: Metrology Definitions and Methods

Two statistical analyses were conducted, based on the type of data: 1) variability of scalar volume measurements, including individual participant performance across test-retest repetitions as well as the performance across algorithms, and 2) comparison of segmentation boundaries relative to reference standard segmentations. The former allows us to compare the performances of these imaging algorithms by measuring agreement of the computed result when the algorithm is held constant as well as when measured by different algorithms, regardless of the similarity in the contours that give rise to the scalar volumes; the latter provides the means by which differing algorithms may be evaluated in terms of the specific segmentation task they are performing which gives rise to the computed scalar volumes.

**Variability of Scalar Volume Measurements**

The models we use assume that the variance is constant across the range of the response. As a first step we determine whether the input data is skewed, which indicates a non-constant variance and perform a transformation suited to the distribution presented so as to achieve constant variance. For this data, the measurement variation was not constant across the range of volumes; it increased with increasing volume measurements as will be illustrated below. Typically, input data is log normal, hence the selection of log transformation as appropriate. To meet the assumptions of the analyses, a log-transformation was applied to volume. As a result, residuals approximately follow a normal distribution which validates the conclusions of the model outputs. Whereas analyses were conducted on the log-scale, data is presented on the original scale, where possible.

Based on the transformed data, we undertook two analyses of volume measurement variability in this study, *repeatability* and *reproducibility* [[22](#_ENREF_22)]. Repeatability refers to variability of measuring tumor volume when repeated measurements are acquired on the same subject under identical or nearly identical conditions. Thus assessment of repeatability approximates the “pure” measurement error of tumor volume measurement. Specifically, we assessed repeatability as the variance of tumor volume measurement when the marker was obtained from repeated imaging of subjects with intentionally short interval so that biological features could be reasonably assumed to have remained unchanged. We assessed repeatability for each of the several participating groups. We further assessed reproducibility, as the variability in tumor volume measurements under the condition where algorithms are not held constant.

We used visual as well as numeric methods to assess variability. Plotting test-retest replications (for repeatability) or pair-wise combinations of algorithms (reproducibility) appear as a straight line of unity in the presence of agreement. Numerically, we denote the measurement of the *j*th algorithm for the *i*th subject at the *kth* replication as *Yijk*, where *j*=1,…,11, *i*=1, …,31, and *k*=1, 2. We used a simple general model , where *Yijk* and *εijk* are the observed value and measurement error and where *μ* is the population mean. is conditional on the mean of infinite replications made on subject *i* by algorithm *j*. Both repeatability and reproducibility were assessed numerically as well as graphically by the Bland and Altman method [[45](#_ENREF_45), [46](#_ENREF_46)]. The method produces an Upper Agreement Limit (*UAL*) and the Lower Agreement Limit (*LAL*) which provides a range within which we expect 95% of the differences between replicate measures of a given algorithm (repeatability) or pair-wise measures of by two algorithms (reproducibility), are expected to lie. Repeatability was represented as differences between each test-retest repetition and plotted against the averages of the two volume measurements.

Based on these analyses, we compute multiple metrics because each provides complementary insight into performance. *RC* is the least significant difference between two repeated measurements on a case taken under the same conditions,

.

The interpretation of *RC* is that the difference between any two normally-distributed measurements on the subject is expected to fall between –*RC* and *RC* for 95% of replicated measurements [[47](#_ENREF_47)]. The 95% tolerance interval for 95% of differences between replicated measurements is , where

and is the thpercentile of the distribution with degrees of freedom.

The within-subject standard of deviation (*wSD*) is estimated as square root of the averaged sample variances across tumors, where the sample variance is computed from the replications for each tumor. This *wSD* assumes that the within-tumor variance is the same across all tumors. The within-subject coefficient of variance (*wCV*) is a relative measure of repeatability, which we calculate as *wSD*/mean and thus is proportional to the magnitude of the tumor’s size.

Concordance correlation coefficient (*CCC)* was computed as in [[48](#_ENREF_48)]. *CCC* is a measure of agreement that is a product of the correlation coefficient, penalized by a bias term that reflects the degree to which the regression line differs from the line of agreement.

Reproducibility was analyzed similarly but instead of the two repetitions, pairwise comparisons were made between algorithms. In this case, the LOA by Bland and Altman provides a range within which we expect 95% of the differences in measurements between two algorithms to lie. The LOA are calculated as ± t(n-1); α/2 *sd*(*Yi1k* – *Yi2k*) (1+1/n), where , where *i*, *j*, *k*, t, and n are as defined above in the repeatability section but now *j* varies pair-wise. Linear Mixed Effects (LME) modeling is used to separate the variability due to subject, algorithm, subject-by-algorithm interaction, and residual. Each of these terms was considered as a random effect in the model. Model assumptions were evaluated with a Q-Q and observed-versus-fitted plots. Use of this model determined the relative contribution to variability by the algorithm as assessed as the sum of the variability to algorithm summed with the subject-by-algorithm interaction as compared with the residual due to other factors in order to inform the QIBA claim by measuring to what extent algorithm versus other variance contributes to overall error.

As done with repeatability, we compute multiple metrics from the reproducibility analysis because each provides complementary insight into performance. The reproducibility coefficient (*RDC*), the interclass correlation coefficient (*ICC*), and the magnitude of variance explained by algorithm versus the residual variance which originates in other factors. Similar to *RC*, *RDC* is calculated as the least significant difference between two measurements taken under different conditions, in our project, by two different algorithms. The *ICC* is a measure of the agreement between the participating groups’ measurements of the CT volumes. The *ICC* is a relative index; it depends on the between-tumor variability. Since the between-tumor variability differs in magnitude for small and large tumors, the *ICC*s of the small and large tumors are not comparable.

In addition to computing the metrics on all tumors, two stratified reproducibility analyses were performed, one by the degree of automation used by the algorithm, and a second by retrospectively dividing into two types: (a) tumors that could be classified as meeting the conditions described in the “Claims” section of the QIBA Profile [[43](#_ENREF_43)], and (b) tumors that did not meet these conditions. Specifically, the claims section of the QIBA profile states that the claims are only applicable “*when the given tumor is measurable (i.e. tumor margins are sufficiently conspicuous and geometrically simple enough to be recognized on all images….) .and the longest in-plane diameter of the tumor is 10 mm or greater*”. Therefore, tumors described as meeting the QIBA Profile were those that were judged to have clearly identified tumor margins; all tumors used in this study exceeded the 10 mm diameter threshold.

**Comparison of Segmentation Boundaries**

Whereas the nature of clinical data makes actual ground truth unavailable, we can form a reference truth if one assumes that those pixels with the highest agreement among participants as being part of the tumor (or not part of the tumor), may collectively be said to be a reference segmentation. We first produced a reference segmentation using the Simultaneous Truth And Performance Level Estimation (STAPLE) method [[49](#_ENREF_49)]. This filter performs a pixel-wise combination of an arbitrary number of input images, where each of them represents a segmentation of the same image, i.e., the segmentations performed by participant algorithms. Each input segmentation is weighted based on its "performance" as estimated by an expectation-maximization algorithm, described in detail in [[50](#_ENREF_50)]. We then compare each individual segmentation result to this reference data, using Sensitivity (*SE*) or true positive rate, calculated as follows. If we define a confusion matrix *C* where *Cuv* is the number of voxels segmented with label *u* while the true label is *v*. For any label *w*, we calculate true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) as:

Typically *SE* is accompanied by Specificity, otherwise known as the true negative rate. However, this quantity has a strong dependence on the size of the field of view which is constant for all participants so we omit reporting this as it is not informative. Rather, the otherwise unused *TP* and *FN* computations are used in the calculation of two additional spatial overlap measures, the Jaccard index [[51](#_ENREF_51)], and Sørensen–Dice coefficients [[52](#_ENREF_52), [53](#_ENREF_53)]:

While at some point it may be evident which is the more important, for this work we compute and present all three types of numeric comparisons, collectively described as “overlap metrics.”

## E.2 Considerations for Performance Assessment of Tumor Volume Change

**Assumption: If the RIDER dataset has 31 lesions, and if most actors have a RC of 15%, then the Profile claim when all imaging procedures are held constant at the two time points must be 21% for a compliance test study with 80% power and 5% type I error rate (one-tailed test).**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Appendix Notes: Test Precision Claim (RC):**

The objective is to test if an actor’s precision is at least as good as the predefined claim value. The precision value in the claim statement of the Profile will be denoted as δ. Let θ denote the actor’s unknown precision. Then the objective is to test the following hypotheses:

 versus .

This will result in a one-sided non-inferiority test. For testing compliance, there are usually only two measurements per case. Then the steps to show compliance are outlined in Table 2.

|  |  |
| --- | --- |
| **STEP** | **DESCRIPTION** |
| 1: Collect the measurements | For each case, calculate the biomarker at time point 1 (denoted Yi1) and at time point 2 (Yi2) where *i* denotes the *i*-th case.  |
| 2. Calculate the change | For each case, calculate the % change in the biomarker: . |
| 3. Calculate the wCV | Over N cases, calculate the within-subject Coefficient of Variation (wCV): .  |
| 4. Estimate the RC | Estimate the Repeatability Coefficient (RC): .  |
| 5. Calculate test statistic and assess compliance | The test statistic T is: , where *claim* is the RC value from the Profile claim statement. Compliance with the claim is shown if , where is the α-th percentile of a chi square distribution with N dfs (for a one-sided test of non-inferiority with α type I error rate) |
| 6. Precision profile | Separate the cases into strata based on covariates known to affect precision. For each stratum, estimate RC. must be < *claim* for each stratum.  |

The profile allows a RC=21%, so this translates into a wCV2 of 57.47%. We can trade wCV2 for Mean Squared Error, where MSE=Var + bias2. Var is the actor’s estimated RC but I took the upper 95% confidence bound on it before I calculated the Bias to be conservative. The sample size in the phantom study should be pretty large.