Study 3A: Inter-method Study with Test-retest Clincal Data: Study Design

Second Challenge

February 2013

Rev 0.3

Document Revisions:

Revision	Revised By	Reason for Update	Date
0.1	Andrew Buckler	Initial version	December 27, 2012
0.2	Andrew Buckler	Revision for team	January 2, 2013
0.3	Andrew Buckler	Updated with file handling and statistical analysis steps graphic and updated module descriptions	February 22, 2013

Table of Contents

1. Introduction

Quantifying changes in lung tumor volume is important for diagnosis, therapy planning and evaluating response to therapy. Computer algorithms have been developed in order to measure such volume changes in clinical settings. The aim of the first QIBA 3A study was to estimate the inter-algorithm variability. The algorithms were applied to FDA acquired CT scans of synthetic lung nodules in anthropomorphic phantoms. Using FDA-supplied physical measurement values as ground truth, we calculate the algorithm measurement accuracy bias and variability. The study was organized as a public "challenge" and consisted of two phases, the pilot and the pivotal one. The objectives of the study where to measure accurate volumes using Computer Tomography (CT) anthropomorphic phantom image data acquired by FDA). The synthetic lung nodules were to be used varied in size (5-40 mm), shape (spherical, elliptical, lobulated, and spiculated), and density (-630, -300, -10, 20, and +100 Hounsfield Units (HU)). Both studies to be used anonymous participants from academic and commercial developers associated with QIBA. The pilot study consisted of 12 participants who measured 97 nodules to perform a feasibility study and a sample size calculation for the pivotal study. The pivotal study consisted of 10 participants who measured 408 nodules. The participants downloaded High Resolution CT images from QI-Bench, an open source software infrastructure that supports the development of quantitative imaging biomarkers. Descriptive statistics and Analysis of Variance (ANOVA) were to be used to test the characteristics of the phantoms and their software-based measurements in terms of volume bias.

This second challenge is undertaken to assess the minimum detectable change of lung lesions imaged on CT using patient datasets as a function of applying heterogeneous algorithms or methods to the same data. The results from this study will broaden the base of data to support the QIBA profile and its descriptions regarding "best practices" for clinical trials and the reduction of measurement variability. The challenge study builds on prior QIBA studies according to the following:

	Variability due to scanner / participant	*Variability due to algorithm / method*
Phantom data	1A, 1C	First 3A challenge
Clinical data	1B	This (second) 3A challenge

1.1. Purpose & Scope

Characterizing performance, such as minimal detectable change, is a prerequisite for biomarker qualification and for establishing limitations when using measurements to determine therapeutic response.

This second 3A challenge uses clinical data sets collected under a no-change condition to determine effects due to differing algorithm or method in the minimum detectable change in size of lung lesions by building on the prior results of QIBA study 1B. CT datasets of 32 non-small cell lung cancer patients scanned twice within 15 minutes and reconstructed as thin transverse slices (publicly available on both TCIA and QI-Bench). One lesion was identified for each patient (32 target lesions). Participants will evaluate each lesion for volumetry. Test-retest repeatability will be evaluated for each participant and across participants by comparing measurements performed on both scans for the same lesion.

Further, this challenge problem increases the type of analysis performed to extend beyond evaluation of the volume results (only), to evaluation of the segmentation boundaries or outlines as well. This is necessary to understand why the results are what they are, and provide important insights for developers and suppliers on the strengths and weaknesses of their algorithms under certain specific clinical conditions as well as providing a basis for optimization using this information.

2. Measures

This second 3A challenge builds on the methodology to be used in prior QIBA studies in two important ways. First, it adopts standardized statistical analysis modules based on the results of the RSNA-sponsored "Metrology Workshop," which had not taken place when study design shall be performed for the prior studies but which is available now. Second, it extends the analysis to an evaluation not only of the computed volume result but as well as pixel-wise analysis of the segmentation objects themselves.

2.1. Metrology Workshop Summary

Definitions were drawn from several sources, including the International Vocabulary of Metrology (VIM; [1]), International Organization for Standardization (ISO; [2]), Clinical and Laboratory Standards Institute (CLSI; [3]), and National Institute of Standards and Technology (NIST; [4]). Throughout the document, the following terms are to be used:

- **Quantity** [VIM, 1.1]: a property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference. This reference can be a measurement unit, e.g., a cubic centimeter (cm^3).
- **Quantity value** [VIM, 1.19]: a number and reference together that express the magnitude of a quantity. For example: the volume of a given tumor, 2.0 cm^3, is a quantity value.
- **Measurement** [VIM, 2.1]: the process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity.
- **Measurand** [VIM, 2.3]: the quantity intended to be measured.
- **Quantitative imaging biomarker (QIB):** an imaging biomarker is quantitative if it meets the following criteria:
  - The difference between two values of the measurand is meaningful.
  - There is a clear definition of zero, in that the ratio of two values of the measurand is meaningful.

  That is, an imaging biomarker is a QIB if the measurand is a ratio variable [5].

  For example: Tumor volume is a QIB when one tumor has a volume of 500 cc and another tumor has a volume of 1500 cc. The following statements have real meaning: 1) the larger tumor is 1000 cc bigger than the smaller tumor; and 2) the larger tumor is 3 times the size of the smaller tumor.

In metrology, it is said that no measured value is complete without an indication regarding its *uncertainty*.

- **Uncertainty** [VIM, 2.26; also called measurement uncertainty, uncertainty of measurement]: a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information to be used.

Uncertainty combines many components. It may derive from the technical performance characteristics of the measure and/or the applicability of the measure to the clinical context for use. Some components of uncertainty arise from systematic effects. Other components of uncertainty arise from random effects. For these reasons, the components of interest contributing to uncertainty are evaluated separately. Significant sources of uncertainty should be identified, and the parameter measuring any of these sources should be stated explicitly. A related term is *variability*.

- **Variability** [NIST, 2.1.1.4]: the tendency of the measurement process to produce slightly different measurements on the same test item, where conditions of measurement are either stable or vary over time, temperature, operators, etc.

Variability in measurements is a general concept. It happens when conditions of measurement are the same, and it is compounded when those conditions differ. Variability in measurements is related to the metrological characteristics of the imaging device when the same test item is measured under stable test conditions. The components of interest contributing to variability are evaluated separately, explicitly identifying each source and stating the parameter being to be used to describe it.

Smaller variability is associated with higher precision (lower standard deviation, i.e., the values are tighter). The number of significant digits in the measurement obtained should reflect the precision. The 'specified conditions' can be, for example, *repeatability conditions of measurement* or *reproducibility conditions of measurement*.

- **Repeatability** [VIM, 2.21; also called measurement repeatability]: measurement precision under a set of *repeatability conditions of measurement*.
- **Repeatability condition of measurement** [VIM, 2.20]: condition of measurement, out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time.
- **Reproducibility** [VIM, 2.25, also called measurement reproducibility]: measurement precision under reproducibility conditions of measurement.
- **Reproducibility condition of measurement** [VIM, 2.24]: condition of measurement, out of a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects.

  Compared with repeatability, reproducibility still requires the same measurement procedure, the same operating conditions, and a short period of time between measurements. It is only location, operator, and/or measuring system that may differ.

When reporting the results of a precision study, a description of the conditions of measurement should be provided. This is especially true if repeatability does not strictly apply. For example, within-center precision can be to be used for a set of conditions that includes different operators (technologists, radiologists), measuring systems, and replicate measurements on the same or similar objects within a single location (center). In that case, between-operator differences and between-instrument differences will contribute to parameters measuring imprecision (e.g., standard deviation, coefficient of variation). Other parameters that might vary in a within-center precision study could include date, time of day for scan, and/or different scanner acquisition settings. Examples of variation in parameters for image analysis include scanner hardware changes, scanner software changes, scan protocol errors, patient motion, patient hydration state, and other sources of variability between patients.

Methodological Considerations: Repeatability

The repeatability of a biomarker measures the ability of that biomarker to detect a change in a single patient. It is a more stringent assessment of performance than when applying the biomarker to a large cohort of patients. It demonstrates the ability of the quantitative imaging biomarker to reliably and repeatably make the same measurement.

Similar repeat measurements may be conducted using multiple phantom scans of varying sizes to obtain a calibration curve (hopefully linear) and an estimate of between scan / within subject variability. The use of phantoms assumes that the actual phantom measurement is known with negligible error. When repeated measurements are conducted on patients, each patient is scanned and measured multiple times and each patient is considered as a randomly selected block of independent and identically distributed (i.i.d.) measurements with mean μ and variance σ^2. The final repeatability variance is estimated from the repeated scans.

2.2. Segmentation Object Analysis

Characterizing the performance of image segmentation approaches has been a persistent challenge [6-8]. Performance analysis is important since segmentation algorithms often have limited accuracy and precision. Interactive drawing of the desired segmentation by domain experts has often been the only acceptable approach, and yet suffers from intra-expert and inter-expert variability. Automated algorithms have been sought in order to remove the variability introduced by experts, but no single methodology for the assessment and validation of such algorithms has yet been widely adopted. The accuracy of segmentations of medical images has been difficult to quantify in the absence of a "ground truth" segmentation for clinical data. Although physical or digital phantoms can help, they have so far been unable to reproduce the full range of imaging and anatomical characteristics observed in clinical data. An

attractive alternative is comparison to a collection of segmentations by experts, but the most appropriate way to compare segmentations has been unclear.

We utilize the Expectation-Maximization algorithm that has been implemented in ITK-Snap for STAPLE for computing a probabilistic estimate of the "ground truth" segmentation from a group of participant segmentations, and a simultaneous measure of the quality of each participant. This approach readily enables the assessment of an automated image segmentation algorithm, and direct comparison of expert and algorithm performance. Additionally, we will explore the utility of a number of other similarity measures, overlap metrics, and pixel-based comparisons on the data. The following table summarizes the range of methods under consideration:

Metric	Purpose	Source	Language	Status (as of 12/13/2012)
STAPLE	To compute a probabilistic estimate of the true segmentation and a measure of the performance level by each segmentation	FDA	MATLAB	testing
STAPLE	Same as above	ITK	C++	implemented
soft STAPLE	Extension of STAPLE to estimate performance from probabilistic segmentations	TBD	TBD	TBD
DICE	Metric evaluation of spatial overlap	ITK	C++	implemented
Vote	Probability map	ITK	C++	implemented
P-Map	Probability map	C. Meyer	Perl	TBD
Jaccard, Rand, DICE, etc.	Pixel-based comparisons	Versus (Peter Bajcsy)	JAVA	TBD

3. Study Design

Our primary hypothesis is that the minimal detectable change in tumor size is smaller than the QIBA profile claim. Moreover, our secondary and tertiary objectives are to evaluate individual segmentation object using pixel-wise indices.

There are three primary actors: the participant, the registrar, and the trusted broker:

1. Individual participant:
    - Method (including any algorithms to be used) included in the imaging test for data and results interpretation must be pre-specified before the study data is analyzed. Participants will be provided a development set for any algorithm tuning, such development set to be comparable to the test set, but without any repeated use of the same data. Lung data is very different from liver, for example.
    - Alteration of the method to better fit the data is generally not acceptable and may invalidate a study.
    - The individual participant or organization will receive back performance data and supporting documentation capable of being incorporated into regulatory filings at its discretion.
2. 3A registrar: handle participant agreements and communications so as to establish and maintain anonymity of participants with respect to the results.

3. Trusted broker:
    - Provides means to archive data sets that may be selectively accessed according to specific clinical indications and that may be mapped to image quality standards that have been described as so-called “acceptable”, “target”, and “ideal”
    - Define services whereby the test set is indirectly accessible via the trusted broker (which means that training data will be accessible each time and only for the test data the user needs contact to the trusted broker).The development set will continued to be available but should be stable whereas test sets may be refreshed with new cases for direct access by interested investigators for testing of new imaging software algorithms or clinical hypotheses.
    - After collection of participant submissions, perform the statistical analysis on the data using open source, standardized, analysis modules.
    - Future investigators will have access to the development set and test sets for additional studies.

3.1. Flow of events for the challenge

The following outlines the procedure to be taken by participants:

- Submit an email to the registrar (non-competing organization) with the signed Participation Agreement and receive an anonymous ID back for identification of results.
- Download and read the 3A Challenge Protocol as posted to the 3A Wiki.
- Download the 3A Challenge data as described in the Protocol. This data will be inclusive of a defined development (e.g., training) set for algorithm adjustment and a test set on which the results would be measured. Data will include images and one location point per target lesion defined by a non-participant.
- Once the development set is to be used by the algorithm to do any parameter tuning, these tuning parameters should be to be used without further modification on the test set (similar to MICCAI liver challenge in 2008). (Note: individual participant integrity is relied on to enforce this policy.)
- Report your results in the required formats, signed by your team leader, to 3A registrar. (Note: this report has to include an method description also.)
- 3A registrar will analyze the reported results as per the Analysis section of this document. 3A registrar will provide Participants with individual analysis of their results. We will publish the results of the evaluation, without publicly identifying individual scores by Participant.

3.2. Study Data

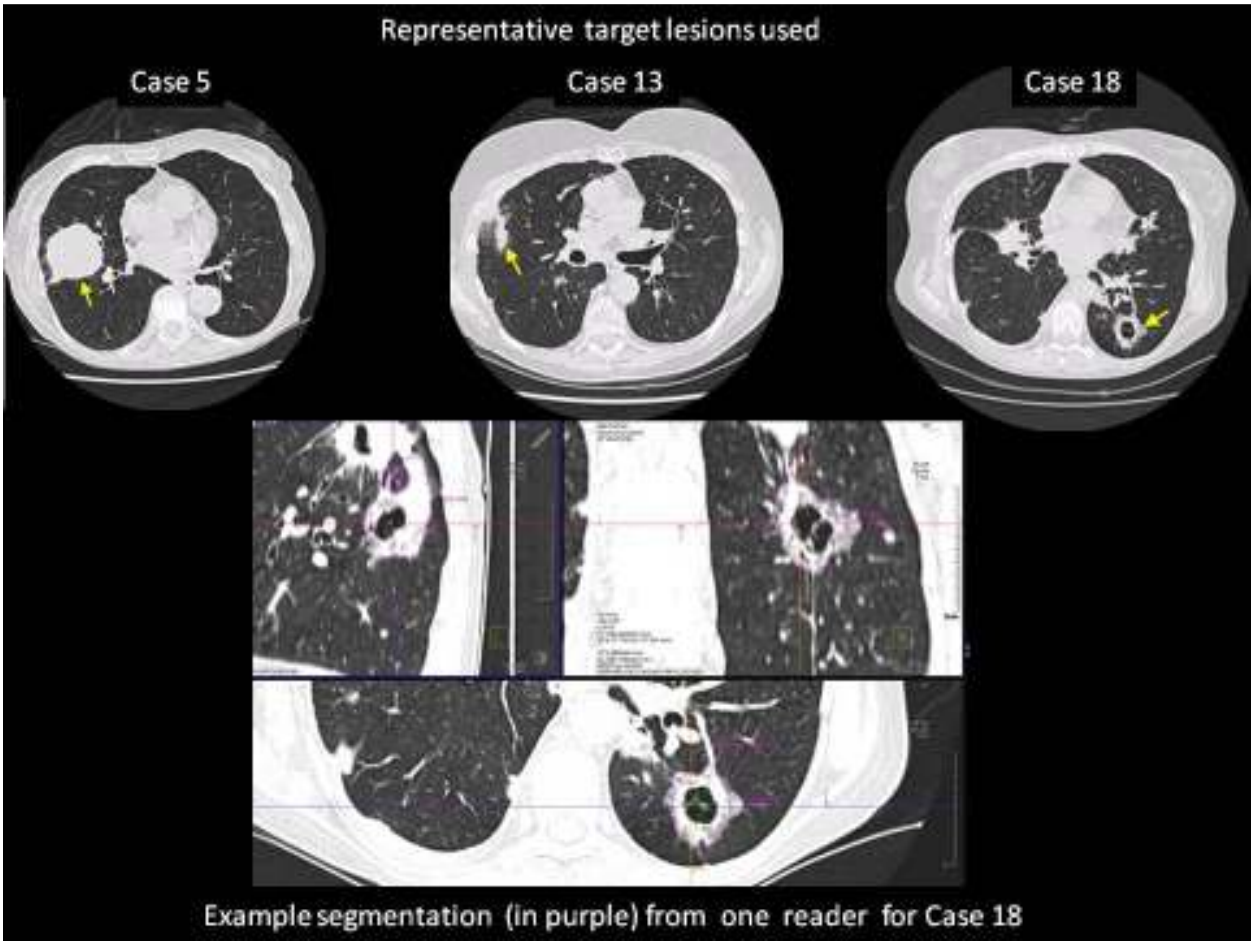
Reference Data Sets will be established and made available to participants, with designated purpose as a “Training” set or a “Test” set. As indicated in Figure 1, an example use is that an challenge study proceeds with a Pilot phase performed using a subset of data, after which a Pivotal phase is run on additional data. The initial Pilot phase includes data partially marked with truth and partially not. The part marked with truth may be utilized for training or optimization purposes and the part not so marked shall be to be used for the pilot test results. After all participants return their pilot results, the full truth markings are made available to the participants thereby creating a larger set that may be to be used for training and optimization prior to running the Pivotal. The Pivotal set is referred to as the Test set, and full truth data is not shared until or unless the community determines that it will not be further to be used for pivotal testing, implying that the test set is refreshed with new data for subsequent pivotal testing.

The cases to be used in this study were drawn from two publicly available image data sets:

- 32 coffee break (no change) cases. These 32 cases were the same ones that were to be used in the original from the original 1B study. They were the “coffee break” cases contributed to the RIDER database from Memorial Sloan Kettering. For that study, each patient was scanned twice within a short period of time (< 15 minutes) on the same scanner and the image data was

reconstructed with thin sections (< 1.5 mm thick). One lesion was selected for each patient for the reading study and subsequent analysis.

- 20 “change cases” also acquired from the RIDER database. In addition to the “no change” cases described above, another 20 cases were extracted from the RIDER database that did have some change between the two time points. These cases were acquired at different time points (typically 3 to 6 months apart) and therefore were expected to have some radiologic change. These cases were also reconstructed with thin sections (< 1.5 mm thick). These cases were to be used as distractor cases so as to reduce participant bias; that is, because they did demonstrate a change in lesion size, participants would not have the expectation that all cases shown to them should have the same size.



3.2.1. Location Coordinates and Ground Truth

Reference Data Sets will be accompanied by location points defined in the context of an indexing scheme. The purpose of this is so as to achieve consistency across participants such that each lesion is identified in the same way. Participants are allowed to select different or multiple seed point(s), provided they utilize the lesion identification scheme provided.

While the individual lesion volumes is not known for this challenge, the test-retest protocol to be used in data collection provides a truth value of zero change across repetitions since there is no actual biological change; hence truth is known to be zero. In practice (and hence the point of the exercise) the measurements made will not be zero hence the interest of the study.

3.3. Instructions to Participants (including specification of read paradigm)

The reading study shall be performed with at least one but participant but optionally more at participant discretion. For the sake of this document it is understood that “participant” may be a human participant (for semi-automated methods) but may also be a computer (for fully automated methods). Participants shall not be told there are both “change “and “no change” cases in this study. They shall only be informed that these were cases with lung lesions and that they should be measured volumetrically.

For each case, one lesion is identified for the participants to measure; this lesion is pre-identified by its exact coordinates (image number and (x,y) coordinates), and communicated to the participants by a “loc” file. Each case consists of two acquisition repetitions (referred to as acqrep 0 and 1 respectively). Cases shall be presented in random order to the participant(s), with a different order for different participants. The actual date and time of the exam shall be hidden from the participant so that they did not know the actual temporal order of each case.

Prior to the commencement of the study, participants may be trained on the reading software. While not required, the “locked-sequential read” paradigm is allowed. When to be used, the following shall be observed:

1. Participants shall read the first acqrep of a given case. Those results shall then be locked with regard to any further editing and no changes shall be made while or after reading the second acqrep.
2. The participants are then presented the second acqrep in such a way that they may refer back to the first acqrep if desired.
3. When the lesion boundary on the second exam was completed, these results were saved and the next case in the session was displayed.

4. Statistical Analysis of this Second Challenge

This second 3A challenge builds on the methodology to be used in prior QIBA studies in two important ways. First, it adopts standardized statistical analysis modules based on the results of the RSNA-sponsored "Metrology Workshop," which had not taken place when study design shall be performed for the prior studies but which is available now. Second, it extends the analysis to an evaluation not only of the computed volume result but as well as pixel-wise analysis of the segmentation objects themselves.

4.1. File Handling and Analysis Steps

Analyses proceeded in three levels, a primary analysis that proceeds based on volume measurements, a secondary analysis which evaluates the segmentation object boundaries across participants at each of the two acquisition repetitions, and a tertiary analysis which evaluates the segmentation object boundaries for each participant but across the two acquisition repetitions. Figure 1 summarizes the steps undertaken for file handling and analysis:

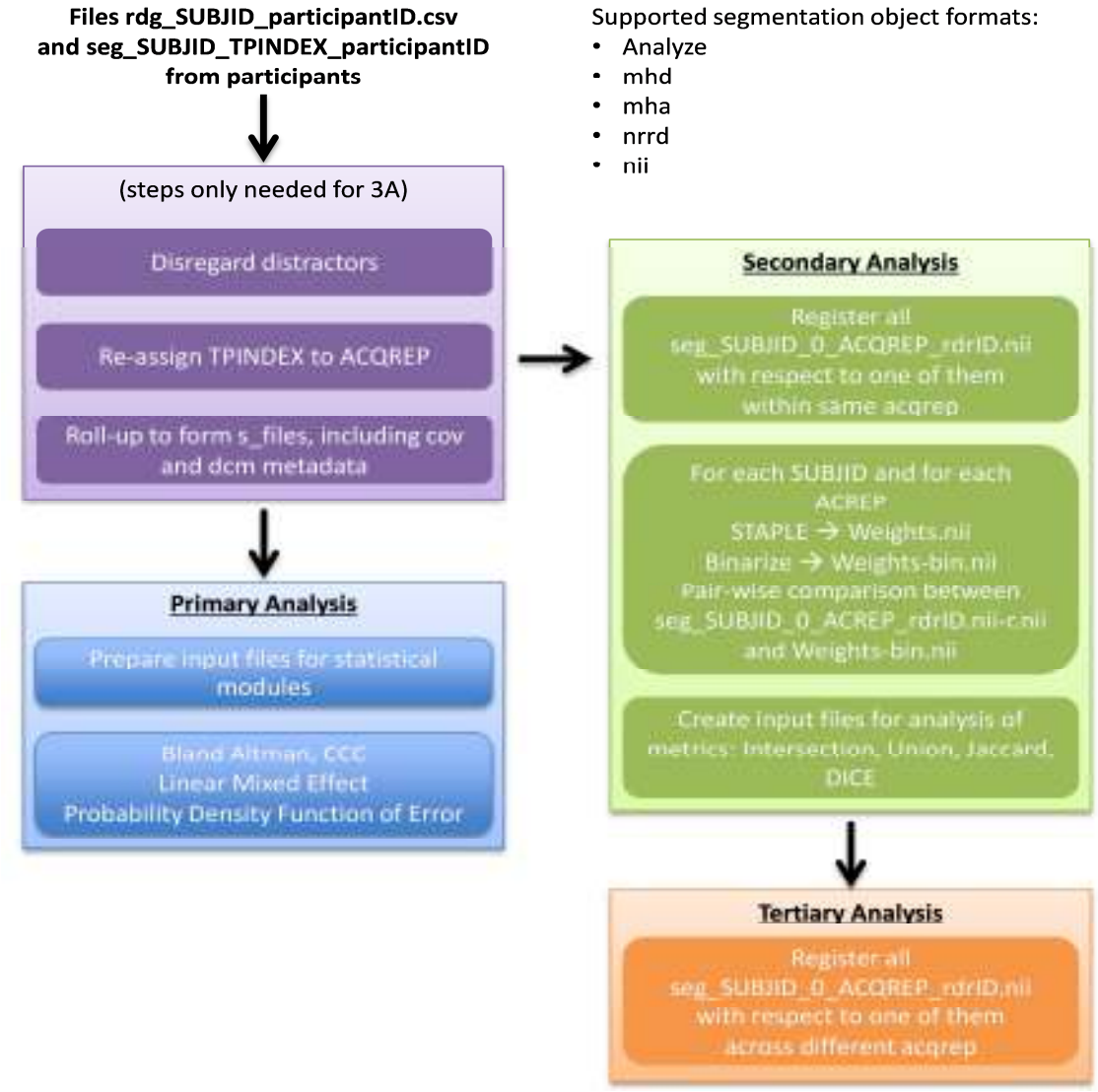


Figure 1: File Handling and Analysis Steps to be Performed

Additional details on the analysis modules are provided in an appendix.

4.2. Data Extraction

The volumes to be used for this Primary Analysis are the ones calculated by participants. The prior work with 1B is used by way of example here.

Various utility modules are utilized to process and classify the data, and to generate the necessary input files. Such files, called ISA-files, are named according to the QI-Bench file name convention and have a format, which complies with the QI-Bench standards. The following is an example of the "rdg" file (s_rdg_1B-Test-retest.csv) where only the first subject of the 1B study is reported.

Table 1: Example Derivation of Reading File Records (aka "rdg")

SUBJID	TPINDEX	ACQREP	TARGET	SEGRDR	SEGTOOL	READINGTYPE	READING
RIDER-1129164940	0	0	TV	1	Python	Volume	44750
RIDER-1129164940	0	0	TV	2	Python	Volume	46520
RIDER-1129164940	0	0	TV	3	Python	Volume	34802
RIDER-1129164940	0	0	TV	4	Python	Volume	47813
RIDER-1129164940	0	0	TV	5	Python	Volume	44770
RIDER-1129164940	0	1	TV	1	Python	Volume	41762
RIDER-1129164940	0	1	TV	2	Python	Volume	45493
RIDER-1129164940	0	1	TV	3	Python	Volume	41687
RIDER-1129164940	0	1	TV	4	Python	Volume	44925
RIDER-1129164940	0	1	TV	5	Python	Volume	43156

The SUBJID is the subject identification number. It is the same Patient's ID value as the image from the TCIA collection. TPINDEX is the time point index, ACQREP is the acquisition repetition number, to indicate a repeat measurement, TARGET refers to the object being measured, in this case it is the tumor volume (TV). SEGRDR is the unique participant identification number, SEGTOOL is the methodology to be used to measure the target, READINGTYPE is the type of lesion to be measured, and READING is the actual measurement.

In addition to the s_rdg_1B-Test-retest.csv, another file is created for the Variability Study. It is the s_dx_1B-Test-retest.csv, which contains, in addition to the SUBJID, TARGET, and X (the average of two readings), the deltaX, which is the error associated to the reading, and the source of the error. The following is an example of the "dx" file, which reports the first subject of the 1B study.

Table 2: Example Derivation of Variability File Records (aka "dx")

SUBJID	TARGET	deltaX	X	SOURCE
RIDER-1129164940	TV	-1769	45635	Inter-participant
RIDER-1129164940	TV	9948	39776	Inter-participant
RIDER-1129164940	TV	-3063	46282	Inter-participant
RIDER-1129164940	TV	-20	44760	Inter-participant
RIDER-1129164940	TV	11717	40661	Inter-participant
RIDER-1129164940	TV	-1294	47166	Inter-participant
RIDER-1129164940	TV	1749	45645	Inter-participant
RIDER-1129164940	TV	-13011	41308	Inter-participant
RIDER-1129164940	TV	-9968	39786	Inter-participant
RIDER-1129164940	TV	3043	46292	Inter-participant
RIDER-1129164940	TV	-3731	43628	Inter-participant
RIDER-1129164940	TV	75	41725	Inter-participant
RIDER-1129164940	TV	-3163	43343	Inter-participant
RIDER-1129164940	TV	-1394	42459	Inter-participant
RIDER-1129164940	TV	3806	43590	Inter-participant
RIDER-1129164940	TV	569	45209	Inter-participant
RIDER-1129164940	TV	2338	44325	Inter-participant
RIDER-1129164940	TV	-3237	43306	Inter-participant
RIDER-1129164940	TV	-1468	42422	Inter-participant
RIDER-1129164940	TV	1769	44040	Inter-participant
RIDER-1129164940	TV	2988	43256	Test-retest
RIDER-1129164940	TV	1026	46006	Test-retest
RIDER-1129164940	TV	-6885	38245	Test-retest
RIDER-1129164940	TV	2889	46369	Test-retest
RIDER-1129164940	TV	1615	43963	Test-retest

4.3. Module for Bland-Altman and Lin's Concordance Correlation Coefficient (CCC)

We utilize two re-usable modules to characterize repeatability and reproducibility. This analysis module produces relatively simple but powerful metrics on input data records representing up to two inter-participant, up to two intra-participant, and up to two test-retest readings to perform relatively popular metrics but without use of a model that produces the most accurate assessments given the latter's ability to account for mixed effects and utilize all availability readings.

Methodology

One of the more popular methods for describing agreement, between- or within-participants, test-retest acquisitions, or in other settings has been promulgated by Bland and Altman in a landmark paper from 1986 [9]. The authors note that "In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analyzed inappropriately, notably by using correlation coefficients. The use of correlation is misleading. An alternative approach, based on graphical techniques and simple calculations, is described, together with the relation between this analysis and the assessment of repeatability." The paper describes a calculation method and a convention regarding how to graphically present the results that we implement here.

Another landmark paper describes the "Concordance Correlation Coefficient" (CCC) that may be compared to correlation coefficients but seeks to avoid a common difficulty with them [10]. The metric seeks to overcome limitations of Pearson correlation coefficients, paired t-tests, and application of least squares analysis. The concordance correlation coefficient is a measure of agreement that is a product of the correlation coefficient that is penalized by a bias term that reflects the degree to which the regression line differs from the line of agreement. The further the regression line is from the line of agreement, the higher the penalty, and the lower the CCC. In has come to be known as Lin's CCC, which we provide here.

Resulting Metrics

The following performance metrics for linearity of quantitative biomarkers are utilized:

- Bland-Altman charts for inter-participant, intra-participant, and test-retest performance, annotated with upper and lower agreement limits.
- Lin's CCC for inter-participant, intra-participant, and test-retest performance.

4.4. Module to compute Linear Mixed Effects Model (LME)

We utilize two re-usable modules to characterize repeatability and reproducibility. This analysis module provides additional insight beyond the relatively simple metrics produced by the Bland-Altman and Lin's CCC methods.

Methodology

This module accepts as input data records representing an arbitrary number of inter-participant, intra-participant, and/or test-retest readings to model multiple fixed and random effects. As such, it is capable of the most accurate assessment due to its ability to account for multiple sources of variability and utilize all availability readings. That said, it is also the most complex and is highly dependent on the appropriateness of model assumptions as well as effect assignment.

Resulting Metrics

The following performance metrics for linearity of quantitative biomarkers are utilized:

- Pareto of effects
- Distribution of model effects, including residuals
- Inter-participant and intra-participant ICC
- QQ-plot from indicating how well the residuals follow a normal distribution

4.5. Module to generate the reference truth segmentation

Methodology

This filter performs a pixelwise combination of an arbitrary number of input images, where each of them represents a segmentation of the same scene (i.e., image). The labeling in the images are weighted relative to each other based on their "performance" as estimated by an expectation-maximization algorithm. In the process, ground truth segmentation is estimated, and the estimated performances of the individual segmentations are relative to this estimated ground truth. The algorithm is based on the binary STAPLE algorithm by Warfield [6]. The multi-label algorithm implemented here is described in detail in [11].

The following is an example of the visualization of segmentation objects:

Resulting Metrics

- Ground truth segmentation.
- Sensitivity and specificity results for each participant as defined by Warfield.
- Visualizations for each participant segmentation result versus the reference truth.

4.6. Module to compute pixel-based comparisons / overlap-based methods

Methodology

The two main overlap measures that are computed are Dice and Jaccard. If we define a confusion matrix C where C_{ij} is the number of voxels segmented with label i while the true label is j. For any label k, we define true positive (TP), true negative (TN), false positive (FP), and false negative (FN) as

$$TP = C_{kk}$$

$$TN = \sum_{\substack{i=1 \\ i \neq k}}^{N} \sum_{\substack{j=1 \\ j \neq k}}^{N} C_{ij}$$

$$FN = \sum_{\substack{i=1 \\ i \neq k}}^{N} C_{ik}$$

$$FP = \sum_{\substack{j=1 \\ j \neq k}}^{N} C_{kj}$$

Resulting Metrics

From these definitions we can define the two spatial overlap measures:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN}$$

and

$$Jaccard = \frac{TP}{TP + FP + FN}$$

These are known as “DICE” and “Jaccard” coefficients. A table of coefficients is produced as well as histograms of values across participants.

Additionally, “intersection” and “union” values are computed, tabularized, and presented as histograms.

5. Reporting

Each participant will be informed (only the anonymized group results with an indication of which member they are). Likewise, the team will produce a publication of the results (to all participants), with authorship representing participants.

At this point, it is possible to apply the study infrastructure to new participants as desired.

Appendix: Endpoints and Investigations of 3A Challenge Studies

Note that this section is maintained from the first 3A study design document for continuity

There are two progressions going on: one from Pilot to Pivotal, and the other, from Primary Investigations to Secondary Investigations. The first derives from the thought that any step may be piloted. Any pilot we do is understood as merely being a miniature of the corresponding pivotal: not different in what's done, only what cases (and the number of cases) we do it on. As such, all steps (including plotting steps) would be done the same way in a pilot as it is on the pivotal. The second progression derives from a step-wise progression from (easier) to (more complex) investigations so that the community learns together over time (Fig. 1).

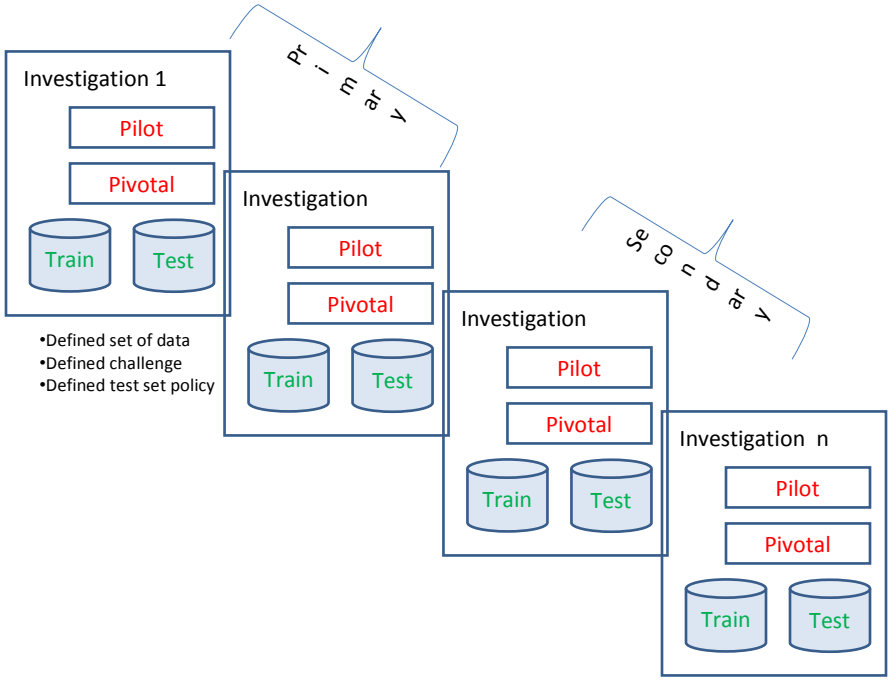


Figure 2: Primary and secondary endpoints: for each challenge study, a Pilot phase is performed using a subset of data, after which a Pivotal phase is run on additional data.

The initial Pilot phase includes data partially marked with truth and partially not. The part marked with truth may be utilized for training or optimization purposes and the part not so marked shall be to be used for the pilot test results. After all participants return their pilot results, the full truth markings are made available to the participants thereby creating a larger set that may be to be used for training and optimization prior to running the Pivotal. The Pivotal set is referred to as the Test set, and full truth data is not shared until or unless the community determines that it will not be further to be used for pivotal testing, implying that the test set is refreshed with new data for subsequent pivotal testing.

References

1. *Joint Committee for Guides in Metrology, "International Vocabulary of Metrology – Basic and General Concepts and Associated Terms. Available from: http://www.nist.gov/pml/div688/grp40/upload/International-Vocabulary-of-Metrology.pdf, accessed 27 November 2011.*
2. *International Organization for Standardization* Available from: http://www.iso.org/iso/home.html, accessed August 1, 2012.
3. *Clinical and Laboratory Standards Institute (CLSI)*. Available from: http://www.clsi.org/, accessed August 1, 2012.
4. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. 1993; Available from: http://physics.nist.gov/Pubs/guidelines/appd.1.html, accessed 27 November 2011.
5. Stevens, S.S., *On the Theory of Scales of Measurement.* Science, 1946. **103**(2684): p. 677-80.
6. Warfield, S.K., K.H. Zou, and W.M. Wells, *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.* IEEE Trans Med Imaging, 2004. **23**(7): p. 903-21.
7. Chalana, V. and Y. Kim, *A methodology for evaluation of boundary detection algorithms on medical images.* IEEE Trans Med Imaging, 1997. **16**(5): p. 642-52.
8. 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.
9. 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.
10. Lin, L.I., *A concordance correlation coefficient to evaluate reproducibility.* Biometrics, 1989. **45**(1): p. 255-68.
11. Rohlfing, T., D.B. Russakoff, and C.R. Maurer, Jr., *Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation.* IEEE Trans Med Imaging, 2004. **23**(8): p. 983-94.

********** END OF DOCUMENT **********