Table of Contents

1. Introduction

X-ray computed tomography (CT) is often an effective imaging technique for assessing therapy response. In clinical practice, qualitative impressions based on nothing more than visual inspection of the images are frequently sufficient for making some clinical management decisions. Quantification becomes helpful when tumor masses change slowly over the course of illness. Many investigators have suggested that quantifying whole tumor volumes could solve many of the limitations of RECIST's current dependence of uni-dimensional diameters on axial slices, and have a major impact on patient management.[1,2] A few studies have shown that volumetry has value.[3] Some reports about the precision[4,5,6] and accuracy[7] of measurement have led to concerns about the risks of confusing variability with medically meaningful changes.

QIBA[8] has constructed a systematic "process map"[9] to qualifying volumetry as a biomarker of response to treatments for a variety of medical conditions, including lung disease. Several trials are now underway to provide a head-to-head comparison between volumetry and RECIST in multi-site, multi-scanner-vendor settings. The QIBA Profile is expected to provide specifications that may be adopted by users as well as equipment developers to meet targeted levels of accuracy and clinical performance in identified settings, both as a correlation to clinical outcomes as well as a comparison to the accepted measure of uni-dimensional diameters.*

One approach to encouraging innovation that has proven productive in many fields is for an organization to announce and administer a public "challenge" whereby a problem statement is given and solutions are solicited from interested parties that "compete" for how well they address the problem statement. The development of image processing algorithms has benefitted from this approach with many organized activities from a number of groups. Some of these groups are organized by industry (e.g., Medical Image Computing and Computer Assisted Intervention or MICCAI[10]), academia (e.g., at Cornell University[11]), or government agencies (e.g., NIST[12]). This workflow is intended to support such challenges.

It is important to note that one of the reasons for doing this 3A study is to meet the need that a biomarker is defined in part by the "class" of tests available for it. That is, it is not defined by a single test or candidate implementation but rather by an aggregated understanding of the results of such tests. As such, it is necessary through this or other means to organize activities to determine how the class performs, irrespective of any candidate that purports to be a member of the class. The corresponding workflow is related to the "Compliance / Proficiency Testing of Candidate Implementations" workflow and it may be that an organization such as NIST can both host challenges as well as serve in the trusted broker role using common infrastructure for these separate but related functions.

In summary, 3A is motivated by the following:

- Changes in malignant nodule volume is important for diagnosis, therapy planning, therapy response evaluation
- Measuring volume changes requires high accuracy in measurement of absolute volume
- Volumes of synthetic nodules may be measured with high accuracy
- Therefore it make sense to use such phantom data (as ground truth) in order to calculate accuracy measurement of algorithms
- The study results could be combined with the QIBA 1A and 1B Group work. This combination will improve the QIBA volumetric CT Profile development.

* It should be noted that RECIST has not been shown to reliably achieve an accurate and precise measurement with a 20% SLD measurement.

We will proceed to add reference clinical data sets, e.g., from Volcano, LIDC and other studies, moving forward wherein at least volume change can be measured by these algorithms, even if ground truth is not available (e.g., no pathologic specimen to compare to)

1.1. Purpose & Scope

The primary and first aim of the study is to estimate inter- and intra-algorithm variability by the volume estimation of synthetic nodules from CT scans of an anthropomorphic phantom (according to the work of the QIBA 1A Group. An inter-algorithm study, in the same way QIBA has been working on inter-reader, inter-scanner, and inter-site. We will also connect the output of this study to the analysis section of QIBA Profile. The aim of the study is not a measure of which of several algorithms provides the best image analysis. Rather, the aim of the study is to gain knowledge to improve QIBA Volumetric CT Profiles and to provide context in which multiple parties have incentives to participate, while avoiding competition and supporting cooperation with a conjoint approach.

Participants include academic and commercial algorithm developers. Industrial vendors may include, for example possible vendors, according to the Volcano 2009 challenge could be: Siemens, Philips, MeVis, Kitware, Definiens, Intio, VIA CAD etc…)

Scope of the study includes the following types of approaches:

- An automatic segmentation algorithm does not require any user intervention (include detection).
- A semi- automatic algorithm needs minimal amount of input from user, e.g., a seed point to initialize the segmentation, then
  - user allowed to edit
  - user does not edit

The first step of the study will be a pilot one. For that pilot study a single seed point will be used to initialize the segmentation and users are not allowed to edit.

Further steps (feature) of the study may use multiple seed point and the user *is* or *is not* allowed to edit. In case the user is allowed to edit then:

Comment [d1]: I agree

Description/Classification of the algorithms: according to the grade of user intervention is needed (for example Volcano'09, A. P. Reeves et al):

- Totally automatic using seed points (no editing beyond setting initial seed)
- Limited parameter adjustment (on less than 15% of the cases)
- Moderate parameter adjustment (on less than 50% of the cases)
- Extensive parameter adjustment (more than 50% of the cases)
- Limited image/boundary modification (on less than 15% of the cases)
- Moderate image/boundary modification (on less than 50% of the cases)
- Extensive image/boundary modification (more than 15% of the cases)

Comment [DEG2]: Isn't this section now wrong? It seems inconsistent with our stated scope if we are not allowing manual editing of detected borders, volume surfaces.

2. Endpoints and Investigations

2.1. Primary Investigations

Study objectives (in priority and sequence order):

1. Characterize Performance of Absolute Volume Estimation where Ground Truth is Known: Results on phantom data, e.g., accuracy and variability, on scans of an anthropomorphic phantom (according to the work of the QIBA 1A Group.[13] (see Dr. Petrick's paper, SPIE 2011). Results on these data utilize the absolute volume primary endpoint.
2. After an initial pilot study (1 above) we will proceed to Characterize Performance of Change Estimation in the Absence of Biological Change for results on clinical data, e.g., minimum detectable change and reproducibility. The study will utilize the same data set as used in the 1B study. Results on these data utilize the percent change in volume primary endpoint.

The primary investigations are accomplished using participant-supervised reads (they are able to train and otherwise prepare readers for the optimal use of the algorithm, as well as have full access to the study data sets and results). (I don't understand the meaning of these sentences)

The QIBA Profile is used to establish targeted levels of performance with means for formal data selection that allows a batch process to be run on data test by a trusted broker that is requested by commercial entities that wish to obtain a certificate of compliance (formal) or simply an assessment of proficiency as measured with respect to the Profile. (Does this refer to 1. or to 2. or to both?)

2.2. Secondary investigations (future plans, needs study design extension in the future)

Extended study for additional measures of algorithm effectiveness:

- MICCAI-like objective criteria of algorithm performance [†]
- Usability or workflow-effectiveness evaluations (e.g., how many corrections, how fast the algorithm ran (time), etc.

Extended study to explore trusted broker scenarios:

- Sequester data and have a black-box wrapped full automated algorithm produce results;
- Support impartial readers to use participant-defined interfaces to their algorithms (allowing whatever training they recommend beforehand but otherwise not have access to the raw imagery prior to the test).

These secondary investigations are expected to utilize the informatics infrastructure but are otherwise not being actively developed at this time.

Extended study to characterize performance of change estimation in the presence of biological change is generally understood to be out of scope for 3A at this time, covered in QIBA's 3B effort. This does not preclude use of the infrastructure for such studies, however, and in fact it is understood as desirable for this to be supported in the future.

2.3. Primary and secondary endpoints

3A (current focus):

1. Investigation (scope of participant agreement)
    - Defined challenge: estimate absolute volumes in phantom data
        - Explicitly indicate experimental factors (primary: analysis sw model. Secondary is acquisitions settings)

Comment [d3]: This paragraph means that we'll expose all the data to participants, which gives them the possibility of manipulating their results if they had a mind to « cheat » or even just inadvertently without ill intent they could use the test data to optimize their algorithms which would skew the results. But we trust the particpants not to do this.

Later, in the « trusted broker » scenarios, this would be enforced since participants don't see the test data.

Comment [d4]: This paragraph can be removed. In part it is not necessary, and in part not accurate.

Comment [d5]: yes

[†] Deng X and Du G. 3D segmentation in the clinic: A grand challenge II – liver tumor segmentation 2008. http://lts08.bigr.nl

- Explicitly indicate descriptive statistics: bias, variance
- Null hypothesis: individual algorithms have bias and/or variance exceeding 15%. (alternative is that both bias and variance of any individual algorithm is less than 15%)

Comment [d6]: do an error propagation model ?

- Policy regarding test data: participants see test data but are asked not to use it in algorithm optimization. Training data is available for any use to support their participation in the challenge that the participant desires (including not using it at all, for example, if they feel they have sufficient data on which their algorithm has been optimized).
- Defined set of data (cases drawn from the FDA CDRH data set):
  - All data associated with the challenged shall comply the QIBA profile.
  - Pilot: for sandbox practice, and to be used in the power study for the pivotal:
    - Training: 5 cases, all lesions per case
    - Test: same 10 cases, all lesions
  - Pivotal: for published results:
    - Training: participants may use all of the 15 pilot cases for algorithm optimization
    - Test: <*number of cases/lesions as determined in the power study as derived based on the pilot*> cases

2. Investigation (next up)
   - Defined challenge: estimate volume change
   - Policy regarding test data:
   - Defined set of data (cases same as was used in 1B):
     - Pilot
     - Pivotal

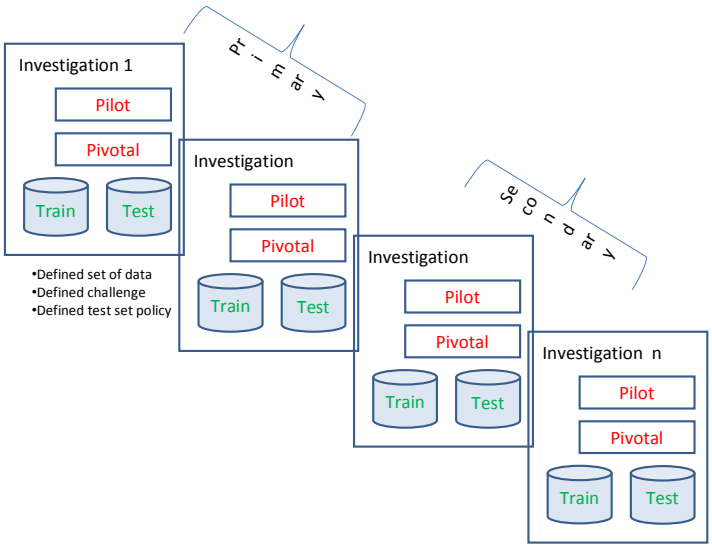

Fig. 1 Primary and secondary endpoints

The primary endpoint depends on the study data utilized. In such cases where there is a single time point per case, the primary endpoint is volume in mm^3. In such cases where there are two time points, the primary endpoint is percent change in volume from the first time point.

There are no secondary endpoints.

3. Study Design

3.1. Data

Reference Data Sets will be established and made available to participants. In the primary investigations, these data comprise the set on which the test is to be run, with the stipulation that these data NOT be used by participants for algorithm training or optimization. (Which means we will not give them a training set?).

Extended studies to explore trusted broker scenarios may subsequently be pursued wherein so-called "training" or "development" sets may be made available for the convenience of participants, but the definitive test is run on sequestered data not visible to the participants. These extended studies are out of scope presently but aspects of the investigational infrastructure are intended to be used for them.

3.2. Seeding and Truthing

Reference Data Sets will be accompanied by seed 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. The first step is to constrain participants to the identified seed point, extended studies may be subsequently pursued wherein participants may be allowed to select different seed point, provided they utilize the lesion identification scheme provided.

Comment [d7]: yes, this is what it means. However, we can change it if we like, in the sense that we could make the reference data sets more granular so as to indicate one subset for training and another for the test. The infrastructure allows both, we just need to decide what we'd like to do.

Comment [DEG8]: We have been using the term pilot phase, pilot or pivotal here – either is OK, but for consistency I suggest pilot.

Truth is known in both of the primary investigations. Phantoms are utilized for the absolute volume estimation study and external measurements are made. The volume change study is performed on scans separated by approximately 15 minutes so as to establish that there is no actual volume 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. Statistics

Statistical measures calculated in these studies include Uncertainty (specifically Bias) and Variability (specifically Variance). Interestingly, we don't include:

Accuracy among these, as the term is formally understood to be a qualitiative one that reflects whether a result is "right" or "wrong" rather than a numeric quantity. As such, we do estimate Bias, but reserve Accuracy for the clinical interpretation of these measurements rather than applying at the level of the measurements themselves.

Likewise, note that Precision and Reliability (comprised of Repeatability and Reproducibility) is not assessed in these studies. The reason for this is we presently don't include repeated measurements and/or use of alternative seed points. It is possible to pursue extended studies that estimate these performance characteristics as well but the present scope precludes them.

3.3.1. Characterizing Performance of Absolute Volume Estimation where Ground Truth is Known

This part of the study utilizes a reference data set comprised of the same cases as used in the 1A study, and extends the statistical analysis that was conducted for that study but in the standard terms as defined above. Specifically, the following parameters are assessed:

- Uncertainty
  - Bias: mean of measured volume minus the physical measurement of the anthropomorphic phantom object. Expressed as percent of actual.

$$Bias = \sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{..})/N$$

where Y_{ij} is the percent difference in volume (i.e. (measured –phantom size) /phantom size*100) in i[th] phantom and measured by j[th] algorithm, $\overline{Y}_{..}$ is the mean of the percent difference across phantoms and algorithms, *N* (= n×k) is the number of observation in the sample set.

- Variability
  - Variance: estimate overall variance in the difference of measured volume from known physical measure or in the difference of two calculated measured volumes in the same tumor in two images (e.g. different factor levels; slice thickness).

$$Mean\ of\ Total\ Variance = \sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{..})^2/(N-1)$$

$$Mean\ Square\ Error\ within\ Algorithms = \sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{i.})^2 / \ (N-k)$$

$$Mean\ Square\ Error\ between\ Algorithms = \sum_{i=1}^{n} k(\overline{Y}_{i.} - \overline{Y}_{..})^2/(k-1)$$

where Y_{ij} is the percent difference (i.e. (measured –phantom size) /phantom size*100)

Comment [M9]: Here are extra parenthesis and an extra k at numerator?

in i^{th} phantom and measured by j^{th} algorithm, $\overline{Y_{..}}$ is the mean of the percent difference across phantoms and algorithms, $\overline{Y_{i.}}$ is the mean of relative bias across algorithms, *N (*= n×k) is the number of observation, n is the number of phantom, and k is the number of algorithm in the sample set.

Comment [M10]: May be redundant with the text below « Bias » formula.

The above is assessed at two levels. First, the group of tests that collectively comprise the so-called acceptable assay methods for the biomarker.[‡] Second, the performance of individual test, in terms of how the individual results compare with the dispersion evident in the group.

1. Perform the methods on the reference data:
   - .1 Analyze statistical variability across the following factors: 1. algorithm type, 2. Image formation factors (exposure, pitch, collimation, slice thickness, reconstruction kernel), 3. Anthropomorphic features (size, morphology, density, and attachment).
     - .1.1 Overall: estimate bias and variance using mean, SD, box-plot (as a more flexible representation than BA) in the difference of measured volume from the physical volume of phantom
     - .1.2 Similar analysis for each factor
     - .1.3 (group discussion needed on whether degree of automation is in scope)
   - .2 Additionally, perform ANOVA or regression analysis to test the variability among algorithms and degree of automations
   - .3 Identify outliers whose bias are greater than 30% and report a summary in characteristics of tumor
2. Assess the performance of each descriptive statistic (in our case, Bias and Variance but more generally to include others such as Accuracy, Precision, Repeatability and Reproducibility) and describe them in a box plot similar to the following example:

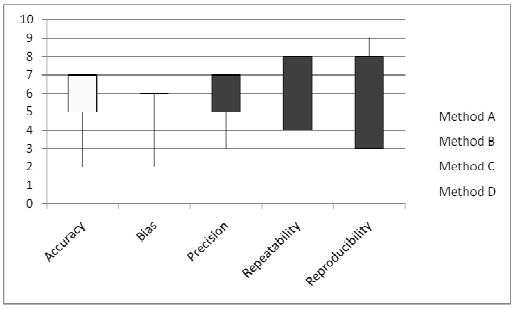

Figure 2: Box plots showing dispersion of participant results for each of the descriptive statisitics selected for the study (Bias and Variance to be used for the first work here, but this example shown extended to include other descriptive statistics also).

3. Select a “group value” for each of the descriptive statistics, e.g., as the mean plus 2 std.
4. For each participant, report their results back to them in the following form:

[‡] For our purposes, since we are gaining experience with this concept, we arbitrarily consider the list of participants to comprise this group. As we gain experience with this method and/or replace it with better ones, we will tighten how we define “acceptable methods” in line with the QIBA Profile specifications.

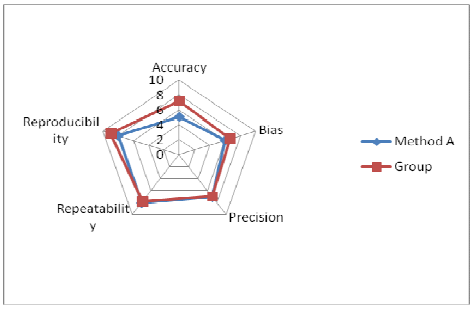

Figure 3: Radar plot showing the "group value" and how one of the individuals compares with it (Bias and Variance to be used for the first work here, but this example shown extended to include other descriptive statistics also). (Which means this plot will be not used for the pilot study)

In a later phase, this analysis may be conducted over the set of available FDA acquisitions, e.g., the 1187 cases presently loaded as a larger reference data set.

3.3.2. Characterizing Performance of Change Estimation in the Absence of Biological Change

This part of the study utilizes a reference data set comprised of the same cases as used in the 1B study, and extends the statistical analysis that was conducted for that study but in the standard terms as defined above. Specifically: the following parameters are assessed:

- Uncertainty
  - Bias: mean of changes in two measured volumes. Expressed as percent of actual.

$$Bias = \sum_{i=1}^{n}\sum_{j=1}^{k}((\frac{X_{ij2} - X_{ij1}}{X_{ij1}} * 100) - \overline{Y_{..}})/N$$

, where X_{ij1} is a volumetric measurement in i[th] phantom and measured by j[th] algorithm at the time point 1, X_{ij2} is percent changes in i[th] phantom and measured by j[th] algorithm at the time point 2, $Y_{ij} = \frac{X_{ij2} - X_{ij1}}{X_{ij1}} * 100$, $\overline{Y_{..}}$ is set to zero as the mean in the absence of biological changes, N=n×k is the number of observation, n is the number of case, and k is the number of algorithm in the sample set.

- Variability
  - Variance: estimate overall variance in the difference of measured volume from known physical measure or in the difference of two calculated measured volumes in the same tumor in two images.

$$Mean\ of\ Total\ Variance = \sum_{i=1}^{n}\sum_{j=1}^{k}((\frac{X_{ij2} - X_{ij1}}{X_{ij1}} * 100) - \overline{Y_{..}})^2/(N-1)$$

$$Mean\ Square\ Error\ within\ Algorithms = \sum_{i=1}^{n}\sum_{j=1}^{k}((\frac{X_{ij2} - X_{ij1}}{X_{ij1}} * 100) - \overline{Y_{i.}})^2/(N-k)$$

Comment [d11]: The plot still applies to the pilot.

One of the confusions I was trying to clarify was that there are two progressions going on : one is from « pilot » to « pivotal », and the other, from « promary investigations » to « secondary investigations. »

I think this version of the document doesn't mention pilot, rather, it discusses a progression of studies with a latent assumption that any step may be piloted.

We can make this more explict if we like the idea of doing pilots.

But I would view any pilot we do 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.

where X_{ij1} is a volumetric measurement in i[th] phantom and measured by j[th] algorithm at the time point 1, X_{ij2} is a percent change in i[th] phantom and measured by j[th] algorithm at the time point 2, $Y_{ij} = \frac{X_{ij2} - X_{ij1}}{X_{ij1}} * 100$, $\overline{Y_{..}}$ is set to zero as the mean in the absence of biological changes, $\overline{Y_{i.}}$ is the mean of percent difference in Y_{ij} across algorithms, *N* (= n×k) is the number of observation, n is the number of case, and k is the number of algorithm in the sample set.

The above is assessed at two levels. First, the group of test that collectively comprises the acceptable assay methods for the biomarker. Second, the performance of individual test, in terms of their membership.

1. The following comparisons of markups can be made:

   .1 Analyze statistical variability across the following factors: 1. Algorithm type, 2. Tumor features (size, morphology, measurable/measurable, attachment, contrast of boundary).

      .1.1 Overall: using mean, SD, box-plot using the difference of two time points where there is no biological differences in tumor

      .1.2 Similar analysis for each factor

   .2 Calculate the precision of the changes in estimated volumetric measures in thresholds of 5%, 10%, 15%, 20%, 25%, and 30%.

      .2.1 Overall: using mean, SD, graph of the precision (%) vs. the thresholds, water-fall plot

      .2.2 Similar analysis for each factor

   .3 Additionally, perform ANOVA or regression analysis to test the variability among algorithms and degree of automations

   .4 Identify outliers whose precision are greater than 30% and report a summary in characteristics of tumor

Subsequently, follow steps 2-4 as defined above using these descriptive statistics.

4. Implementation of the challenge studies (general?)

The following outlines the procedure to be taken by participants:

- Submit an application to participate to the trusted registrar (non-competing organization) and sign the Participation Agreement
- 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 seed point per target lesion defined by a non-participant. (Note: In the pilot study we don't have such a training set, only a test set?)
- Once the development set is used by the algorithm to do any parameter tuning, these tuning parameters should be used without further modification on the test set (similar to MICCAI liver challenge in 2008). (Note: in the pilot phase, 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 algorithm 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

Comment [d12]: Not finished

Comment [d13]: I think so… this is what Hubert meant by this section I think.

Comment [d14]: define

Comment [DEG15]: Perhaps include the PA as an appendix?

Comment [d16]: define

Comment [d17]: define

Comment [d18]: define

Comment [d19]: See comment earlier in the document. We could have separate training and test, it just means more reference data sets. If we want to do this as a team, we can do so. We need to make the decision and then go through this document to make sure we're clear about it.

Comment [d20]: define

Comment [d21]: yes, that is what's meant by 1.1 below.

of the evaluation, without publicly identifying individual scores by Participant.

4.1. Flow of events for each challenge study

In this case there are two primary actors: the participant, and the honest broker:

1. Individual participant:
   .1 Algorithms included in the imaging test (is that an 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 algorithm to better fit the data is generally not acceptable and may invalidate a study.
   .2 The individual participant or organization needs to receive back performance data and supporting documentation capable of being incorporated into regulatory filings at its discretion.
2. 3A registrar:
   .1 The honest broker needs 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”
   .2 It needs to produce documentation regarding results inclusive of a charge-back mechanism to recover operational costs.
3. 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. Future investigators will have access to the development set and test sets for additional studies. *(Why only the test sets? Is that valued and valid for the anthropomorphic phantom data also?)*
4. 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? Is that a future development of the study or we will use the same scenario for the pilot study also?)

Comment [DEG22]: This should not be allowed. If 3D editing is needed, that may be a measure of robustness of the semi-automated or automated segmentation algorithm.

Comment [d23]: This language comes from the trusted broker idea, which we’re not doing now. Part of it we do need the registrar to do, but not all of what is written. It can be simplified for our primary investigations and then expanded later for the trusted broker.

Comment [d24]: All sets would be refreshed over time.

Comment [d25]: Similar to d22, we want part of this language but should simplify it to remove concept of trusted broker for now.

4.2. Results

Each participant has to be informed (only the annonymized 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 initial participants. (Initial participants? A person and the corresponding institution)

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

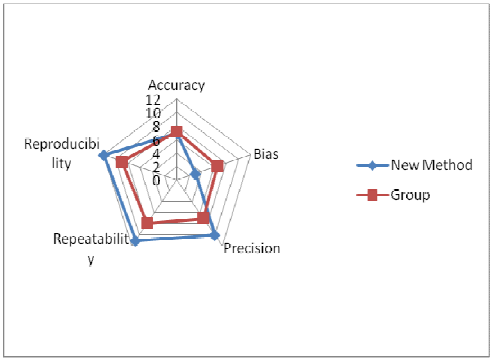

Fig. 4

5. Definitions

- **Uncertainty(2)*:** A value, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurement, composed of uncertainty from both random and systematic error. Random error contributes to reliability, whereas systematic error contributes to validity (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1250265).
    - **Bias:** A quantitative term describing the difference between the average of measurements made on the same object and its true value. In particular, for a measurement laboratory, bias is the difference (generally unknown) between a laboratory's average value (over time) for a test item and the average that would be achieved by the reference laboratory if it undertook the same measurements on the same test item (http://www.itl.nist.gov/div898/handbook/mpc/section1/mpc113.htm).
- **Precision:** Closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions (http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf).
- **Reliability:** *T*he extent to which an experiment, test, or measuring procedure yields the same results on repeated trials (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1250265/).
    - **Repeatability(2)*:** Closeness of the agreement between the results of successive measurements of the same measure and carried out under the same conditions of measurement (http://physics.nist.gov/Pubs/guidelines/appd.1.html).
    - **Reproducibility (2)*:** Closeness of the agreement between the results of measurements of the same measured carried out under changed conditions of measurement (http://physics.nist.gov/Pubs/guidelines/appd.1.html).

- **Variability:** 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. (http://www.itl.nist.gov/div898/handbook/index.htm).
    - **Variance:** the quantity defined as

$$s^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 / n - 1$$

where $\bar{Y}$ is the mean of the data, n is number of observations in the sample set. (http://www.itl.nist.gov/div898/handbook/eda/section3/eda356.htm).

- **Bias:** see above.

6. References

[1] Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced disease. Disease 1976; 38: 388-394.

[2] Quivey JM, Castro JR, Chen GT, Moss A, Marks WM. Computerized tomography in the quantitative assessment of tumour response. Br J Disease Suppl 1980; 4:30-34.

[3] Munzenrider JE, Pilepich M, Rene-Ferrero JB, Tchakarova I, Carter BL. Use of body scanner in radiotherapy treatment planning. Disease 1977; 40:170-179.

[4] Petrou M, Quint LE, Nan B, Baker LH. Pulmonary nodule volumetric measurement variability as a function of CT slice thickness and nodule morphology. Am J Radiol 2007; 188:306-312.

[5] Bogot NR, Kazerooni EA, Kelly AM, Quint LE, Desjardins B, Nan B. Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. Acad Radiol 2005; 12:948–956.

[6] Erasmus JJ, Gladish GW, Broemeling L, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: Implications for assessment of tumor response. J Clin Oncol 2003; 21:2574–2582.

[7] Winer-Muram HT, Jennings SG, Meyer CA, et al. Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations. Radiology 2003; 229:184-194.

[8] Buckler AJ, Mozley PD, Schwartz L, et al. Volumetric CT in lung disease: An example for the qualification of imaging as a biomarker. Acad Radiol 2010; 17:107-115.

[9] Radiological Society of North America. http://qibawiki.rsna.org/index.php?title=Main_Page, accessed 07 Sep 2009.

[10] http://www.grand-challenge.org/index.php/Main_Page, accessed 23 December 2010.

[11] http://www.preventcancer.org/uploadedFiles/Education/Conferences,_Workshops,_and_Educational_Programs/Day-2-Friday-1100-AM-Tony-Reeves.pdf, accessed 23 December 2010.

[12] http://www.nist.gov/itl/iad/dmg/biochangechallenge.cfm, accessed 23 December 2010.

[13] Petrick NP, Kim HJ, Clunie D, Borradaile K, Ford R, Zeng R, Gavrieldes MA, McNitt-Gray MF, Fenimore C, Lu J, Zhao B, Buckler AJ. Evaluation of 1D, 2D and 3D nodule size estimation by radiologists for spherical and non-spherical nodules through CT thoracic phantom imaging, SPIE, February 2011.