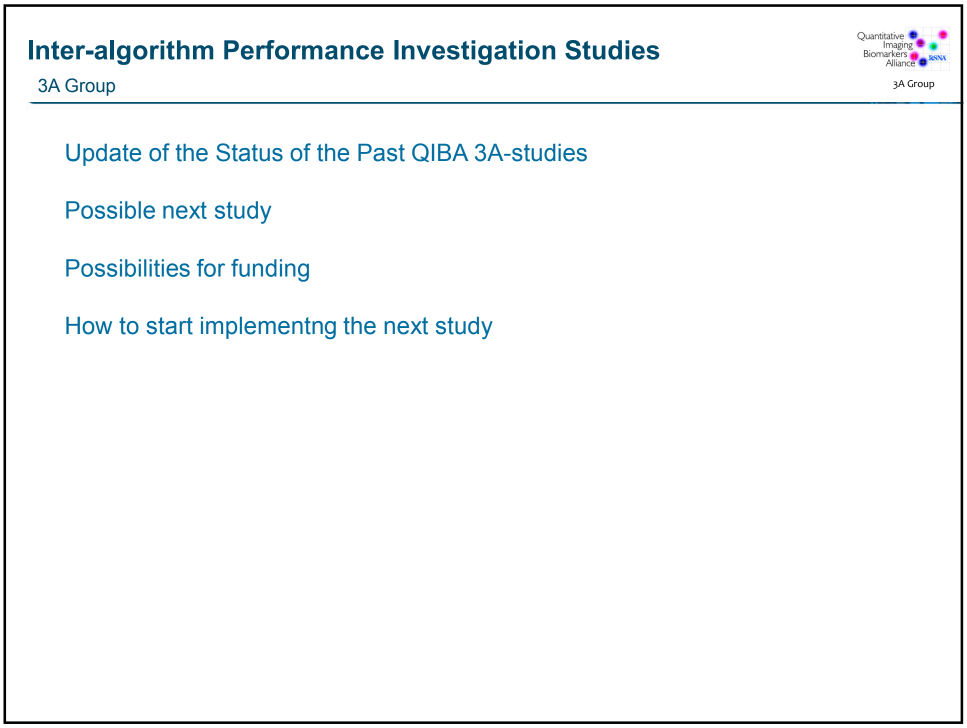


Inter-algorithm Performance Investigation Studies

3A Group

Quantitative Imaging Biomarkers Alliance RSNA

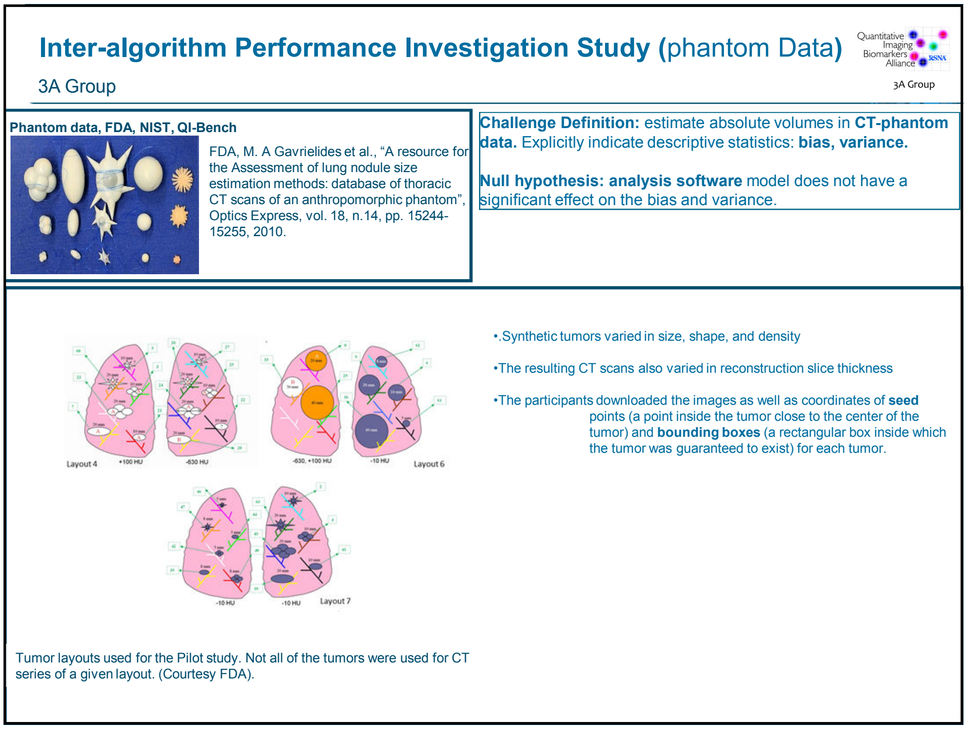
3A Group

Update of the Status of the Past QIBA 3A-studies

Possible next study

Possibilities for funding

How to start implementng the next study



Inter-algorithm Performance Investigation Study (phantom Data)

3A Group

Quantitative Imaging Biomarkers Alliance RSNA

3A Group

Phantom data, FDA, NIST, QI-Bench

FDA, M. A Gavrielides et al., “A resource for the Assessment of lung nodule size estimation methods: database of thoracic CT scans of an anthropomorphic phantom”, Optics Express, vol. 18, n.14, pp. 15244-15255, 2010.

Challenge Definition: estimate absolute volumes in **CT-phantom data.** Explicitly indicate descriptive statistics: **bias, variance.**

Null hypothesis: analysis software model does not have a significant effect on the bias and variance.

Tumor layouts used for the Pilot study. Not all of the tumors were used for CT series of a given layout. (Courtesy FDA).

- .Synthetic tumors varied in size, shape, and density
- The resulting CT scans also varied in reconstruction slice thickness
- The participants downloaded the images as well as coordinates of **seed** points (a point inside the tumor close to the center of the tumor) and **bounding boxes** (a rectangular box inside which the tumor was guaranteed to exist) for each tumor.

Inter-algorithm Performance Investigation Study (phantom Data)

3A Group

Quantitative Imaging Biomarkers Alliance RSNA 3A Group

Results

Descriptive statistics and analysis of variance (ANOVA) were used to test the software-based measurements of phantom volumes in terms of volume bias and variability **(Kim Grace)**.

We studied both the entire set of phantom data, which varied over *size, density, shape, and CT slice thickness,* and also a subset of data containing only those phantoms that met the requirements of the QIBA CT Profile (thin slice ≤ 2.5 mm, size ≥ 10 mm, and solid tumor with excluding density of -630 HU).

We calculated both absolute mean percent error (all measurements > 0) and volume bias, measured as mean percent error (values can be positive or negative), for the entire set and for the subset.

Variation across the participants and all other tumor characteristics are given.

The effects of nodule size, shape, and density, and CT slice thickness were shown to have a statistically significant effect on nodule volume accuracy with p-values< 0.001.

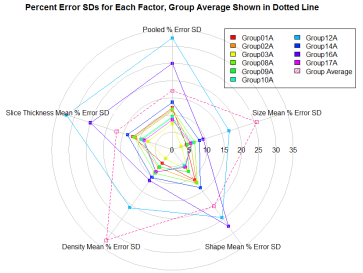
Inter-algorithm Performance Investigation Study (phantom Data)

3A Group

Quantitative Imaging Biomarkers Alliance RSNA 3A Group

Study Results (representative):

10 participants who measured 408 nodules



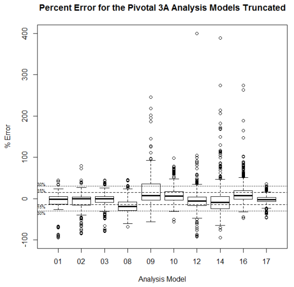


Figure 1 Percent Error for all Participants: the standard deviation from pooled data for all 10 participants are shown by the dotted pink polygon. The pooled standard deviation of each 10 participant is shown by the different colors with a polygon.

Figure 2 Box-whisker plot representing the distribution of the percent error in volume measurements. The mid-bold line indicates the median. The upper and lower lines of box represents 25% and 75% tile in the percent errors. The thicker dashed lines represent ±15%, and the smaller dotted lines show the location of ±30%.

Inter-algorithm Performance Investigation Study (phantom Data)

3A Group

Quantitative Imaging Biomarkers Alliance RSNA 3A Group

Submitted for publication in Academic Radiology. Status: Revision

Title: Algorithm variability in the estimation of lung nodule volume from phantom CT scans: results of the QIBA 3A public challenge.

Maria Athelogou[1], Hyun J Kim[2], Alden Dima[3], Ganesh Saiprasad[3], Adele Peskin[3], Hubert Beaumont[4], Estanislao Oubel[4], Dirk Colditz[1], Marios A Gavrielides[5], Nicholas Petrick[6], Yongqiang Tan[7], Binsheng Zhao[7], an-Martin Kuhnigk[8], Jan Hendrik Moltz[8], Guillaume Orieux[9], Robert J. Gillies[10], Yuhua Gu[10], Ninad Mantri[11], Gregory Goldmacher[11], Luduan Zhang[12], Emilio Vega[13], Michael Bloom[13], Rudresh Jarecha[14], Grzegorz Soza[15], Christian Tietjen[15], Tomoyuki Takeguchi[16], Hitoshi Yamagata[16], Sam Peterson[17], Osama Masoud[17], Andrew J. Buckler[18]

[1]Definiens AG, [2]UCLA, [3]NIST, [4]MEDIAN Technologies, [5]FDA, [6]FDA/CDRH/OSEL, [7]Columbia University Medical Center, [8]Fraunhofer MEVIS Institute for Medical Image Computing, [9]MScGE Healthcare, [10]MScGE Healthcare, [11]ICON Medical Imaging, [12]INTIO, Inc., [13]NYU Langone Medical Center, [14]Perceptive Informatics, [15]Siemens AG, [16]Toshiba Corporation, [17]Vital Images, Inc., [18]Elucid BioImaging, Inc.

Inter-algorithm Performance Investigation Study (clinical data)

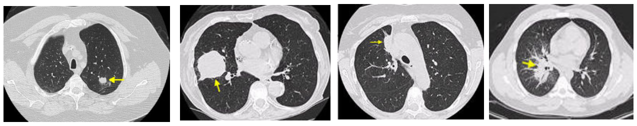
3A Group

Quantitative Imaging Biomarkers Alliance RSNA 3A Group

Challenge Definition: estimate absolute volumes in **CT- clinical data.** Explicitly indicate descriptive statistics: **bias, variance.**

Null hypothesis: analysis software model does not have a significant effect on the bias and variance.

- 41 lung cancer test-retest cases were analyzed by



12 participants in a multi-method study of algorithm performance on the segmentation of clinical CT scans.

- *GE Healthcare*
- *ICON Medical Imaging*
- *KEOSYS*
- *MEDIAN Technologies*
- *Medical University of South Carolina*
- *Mirada Medical*
- *Perceptive Informatics*
- *Fraunhofer MEVIS*
- *Siemens AG*
- *UCLA*
- *University of Michigan*
- *Vital Images*

Inter-algorithm Performance Investigation Study (clinical data)

3A Group

Challenge Definition: estimate absolute volumes in **CT- clinical data.** Explicitly indicate descriptive statistics: **bias, variance.**

Null hypothesis: analysis software model does not have a significant effect on the bias and variance.

- We evaluated variability of scalar volume measurements, in terms of repeatability (individual participant performance across test-retest repetitions)
- reproducibility (performance across participants).
- We also compared segmentation boundaries relative to reference standard segmentations.
- An important outcome of this work is the set of metrics used to define performance for clinical CT data, needed in order to use volume change as a biomarker. These metrics will form a basis for future determination of compliance with the QIBA Profile (**Andrew Buckler**).

Inter-algorithm Performance Investigation Study (clinical data)

3A Group

We evaluated variability of scalar volume measurements, in terms of repeatability and reproducibility. We also compared segmentation boundaries relative to reference standard segmentations. An *important outcome of this work is the set of metrics* used to define performance for clinical CT data, needed in order to use volume change as a biomarker. *These metrics will form a basis for future determination of compliance with the QIBA Profile.*
Repeatability within algorithms is reported in terms of repeatability coefficients **(RC)**, ranging from .06 log (mm^3) (best performing) to 1.5 log (mm^3) (least performing), with corresponding within-subject coefficients of variation of 2.1% to 54% respectively. *Reproducibility* across algorithms is reported in terms of *reproducibility coefficient* **(RDC)**, reported as 0.37 log (mm^3), or about 14% across all tumor sets. *Variability in test-retest measurements* is smaller for a subset of tumors that meet the measurability criterion defined in the QIBA Profile; repeatability of the entire set of tumors is approximately 1.5 times higher than for the subset. *Variability of smaller tumor volumes* was lower without human editing of algorithm measurements, although larger tumors benefitted by editing the algorithm results.
Linear mixed effects modeling led to the conclusion that no more than two-thirds of the overall QIBA Profile variability claim of the system as a whole results from the analysis software (or less if conditions such as the scanner settings are not held constant). Detailed *overlap metrics* as well as *reference segmentations* were provided to participants for their use in optimizing the performance of their methods.

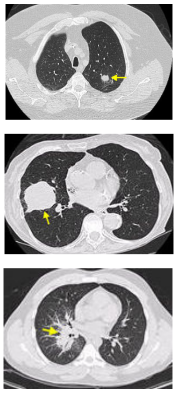
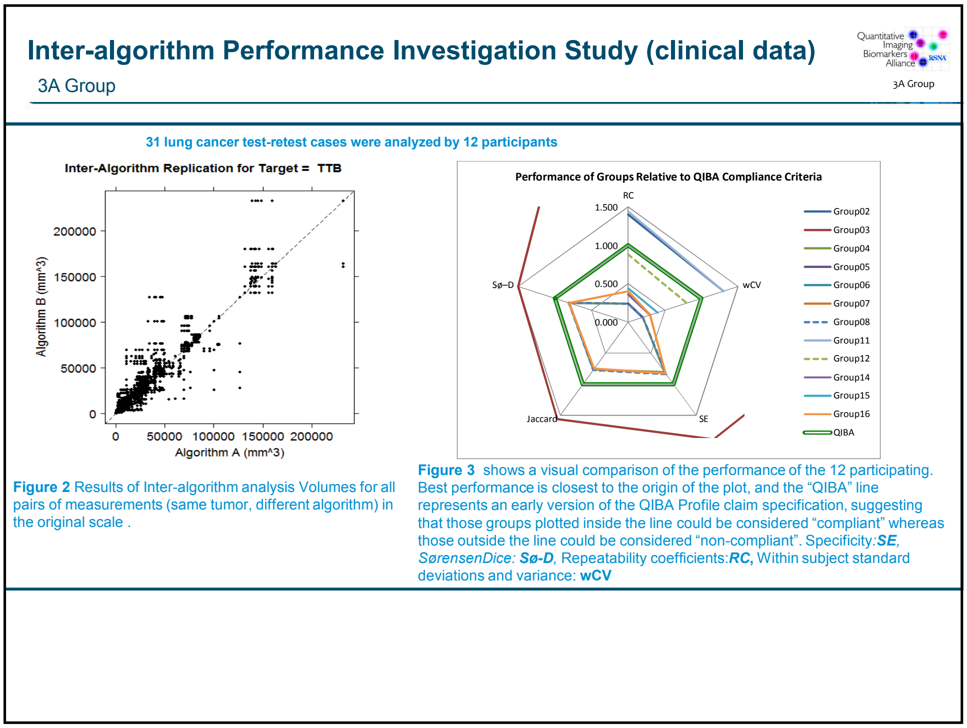


Figure 1: Examples of tumors used



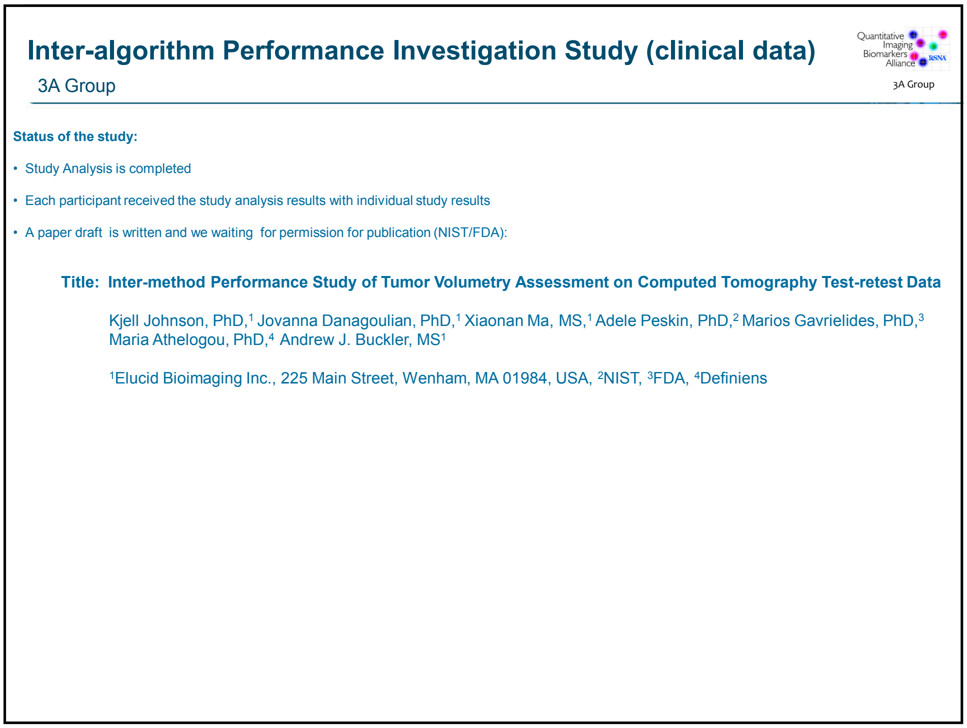
Inter-algorithm Performance Investigation Study (clinical data)

3A Group

31 lung cancer test-retest cases were analyzed by 12 participants

Figure 2 Results of Inter-algorithm analysis Volumes for all pairs of measurements (same tumor, different algorithm) in the original scale .

Figure 3 shows a visual comparison of the performance of the 12 participating. Best performance is closest to the origin of the plot, and the "QIBA" line represents an early version of the QIBA Profile claim specification, suggesting that those groups plotted inside the line could be considered "compliant" whereas those outside the line could be considered "non-compliant". Specificity:***SE***, *SørensenDice:* ***Sø-D****,* Repeatability coefficients:***RC*****,** Within subject standard deviations and variance: **wCV**


Quantitative Imaging Biomarkers Alliance
RSNA
3A Group

Inter-algorithm Performance Investigation Study (clinical data)

3A Group

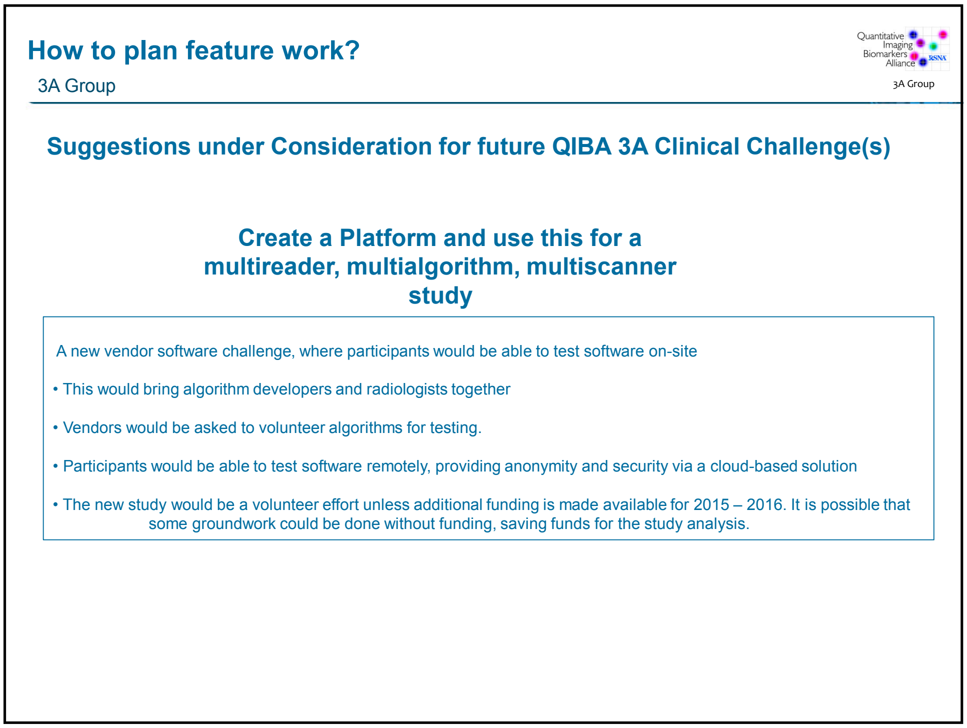
Status of the study:

- Study Analysis is completed
- Each participant received the study analysis results with individual study results
- A paper draft is written and we waiting for permission for publication (NIST/FDA):

Title: Inter-method Performance Study of Tumor Volumetry Assessment on Computed Tomography Test-retest Data

Kjell Johnson, PhD,[1] Jovanna Danagoulian, PhD,[1] Xiaonan Ma, MS,[1] Adele Peskin, PhD,[2] Marios Gavrielides, PhD,[3] Maria Athelogou, PhD,[4] Andrew J. Buckler, MS[1]

[1]Elucid Bioimaging Inc., 225 Main Street, Wenham, MA 01984, USA, [2]NIST, [3]FDA, [4]Definiens



How to plan feature work?

3A Group

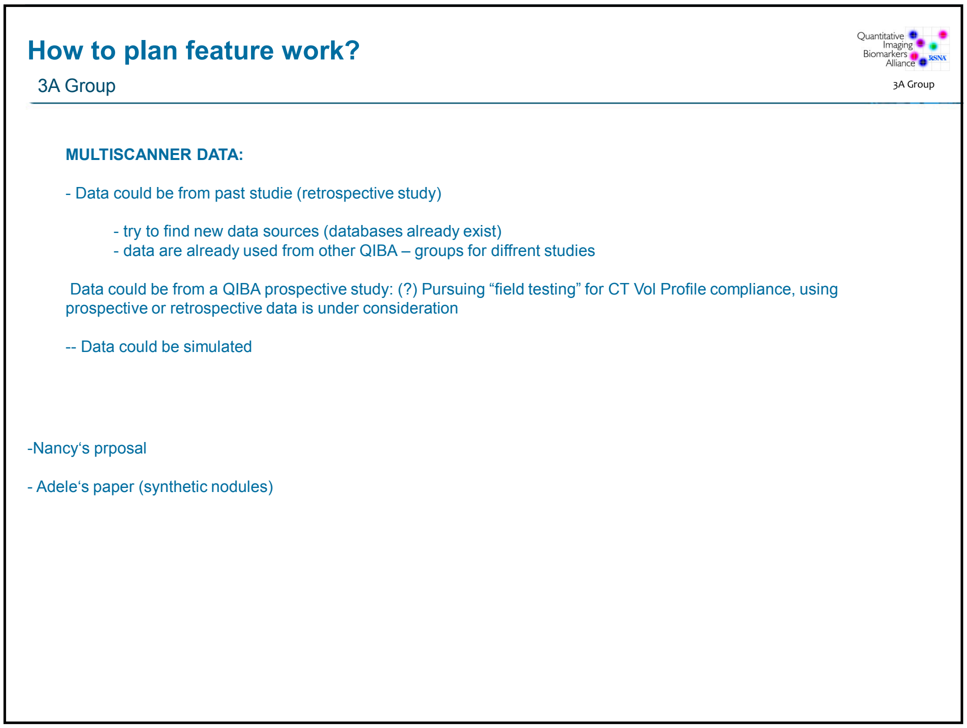
3A Group

Suggestions under Consideration for future QIBA 3A Clinical Challenge(s)

Create a Platform and use this for a multireader, multialgorithm, multiscanner study

A new vendor software challenge, where participants would be able to test software on-site

- This would bring algorithm developers and radiologists together
- Vendors would be asked to volunteer algorithms for testing.
- Participants would be able to test software remotely, providing anonymity and security via a cloud-based solution
- The new study would be a volunteer effort unless additional funding is made available for 2015 – 2016. It is possible that some groundwork could be done without funding, saving funds for the study analysis.



How to plan feature work?

3A Group

3A Group

MULTISCANNER DATA:

- Data could be from past studie (retrospective study)

  - try to find new data sources (databases already exist)
  - data are already used from other QIBA – groups for diffrent studies

Data could be from a QIBA prospective study: (?) Pursuing “field testing” for CT Vol Profile compliance, using prospective or retrospective data is under consideration

-- Data could be simulated

-Nancy‘s prposal

- Adele‘s paper (synthetic nodules)

How to plan feature work?

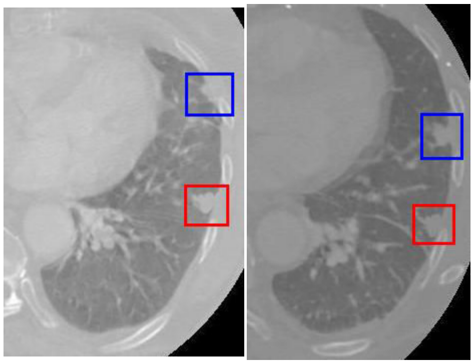
3A Group

Especially: Simulated data could be generated within QIBA:

- Dr. Alele Peskin (NIST) is already familiar with such simulations and (DUKE/FDA Team)

Modeling Clinical Tumors to Create Reference Data for Tumor Volume Measurement

Adele P. Peskin1 and Alden A. Dima2
1 NIST, Boulder, CO 80305
2 NIST, Gaithersburg, MD 20899



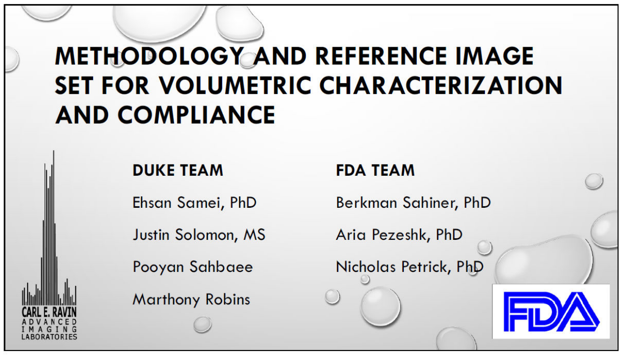
Two time points; the clinical tumors boxed in red, the synthetic tumors in blue, modelled after the tumor at time point 2. Tumor is increased in size by 30 % for the second time point.

How to plan feature work?

3A Group

Especially: Simulated data could be generated within QIBA:

DUKE/FDA Team:

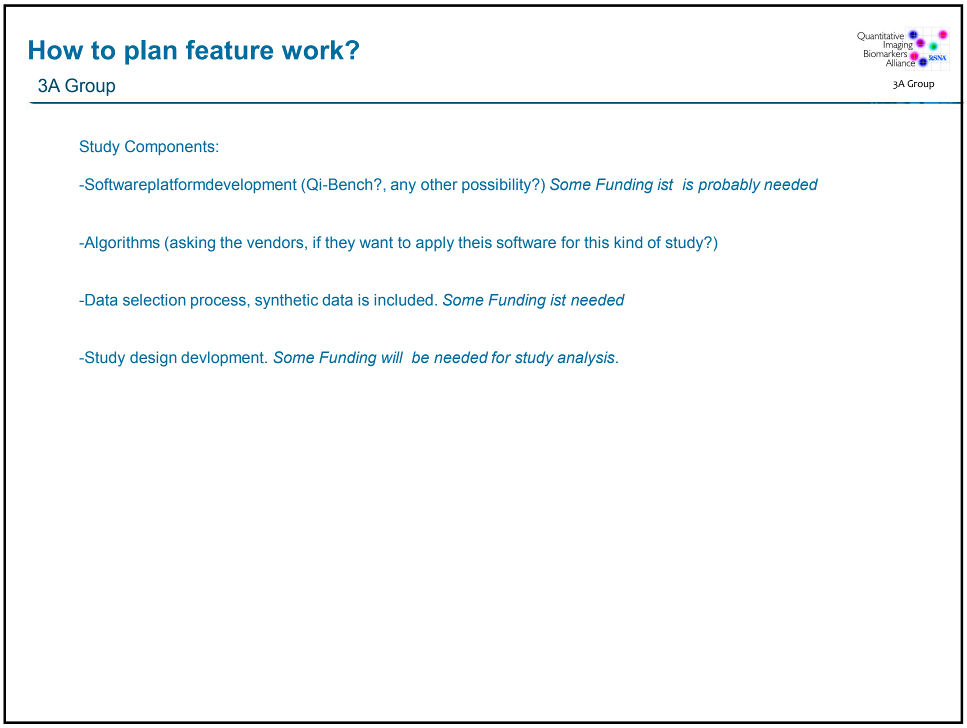


METHODOLOGY AND REFERENCE IMAGE SET FOR VOLUMETRIC CHARACTERIZATION AND COMPLIANCE

DUKE TEAM	FDA TEAM
Ehsan Samei, PhD	Berkman Sahiner, PhD
Justin Solomon, MS	Aria Pezeshk, PhD
Pooyan Sahbaee	Nicholas Petrick, PhD
Marthony Robins	

CARL E. RAVIN ADVANCED IMAGING LABORATORIES

FDA



How to plan feature work?

3A Group

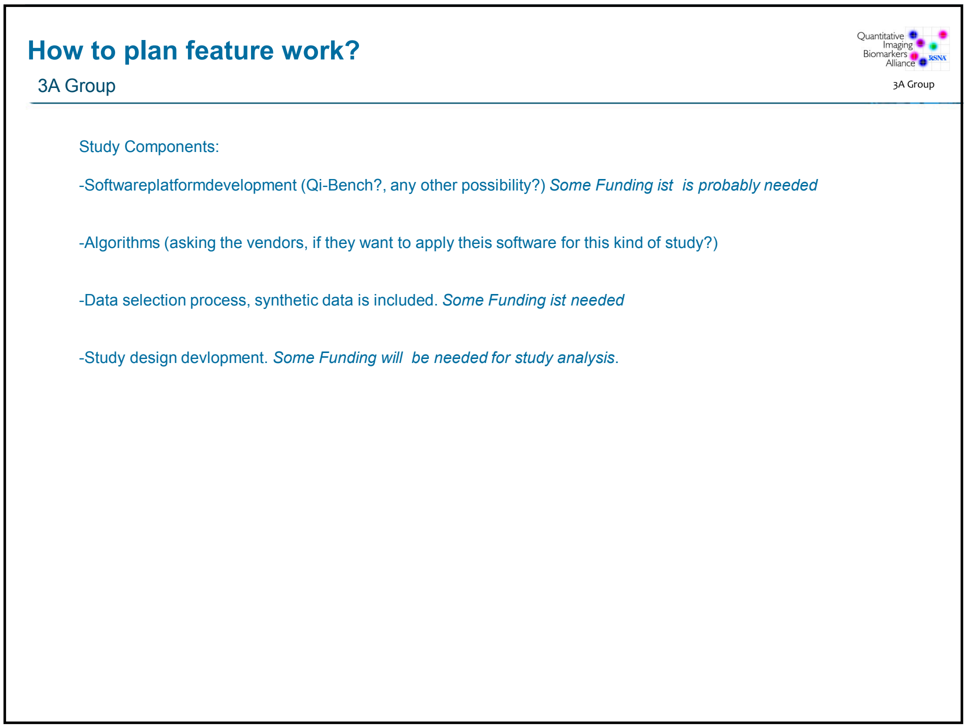
Study Components:

-Softwareplatformdevelopment (Qi-Bench?, any other possibility?) *Some Funding ist is probably needed*

-Algorithms (asking the vendors, if they want to apply theis software for this kind of study?)

-Data selection process, synthetic data is included. *Some Funding ist needed*

-Study design devlopment. *Some Funding will be needed for study analysis*.



How to plan feature work?

3A Group

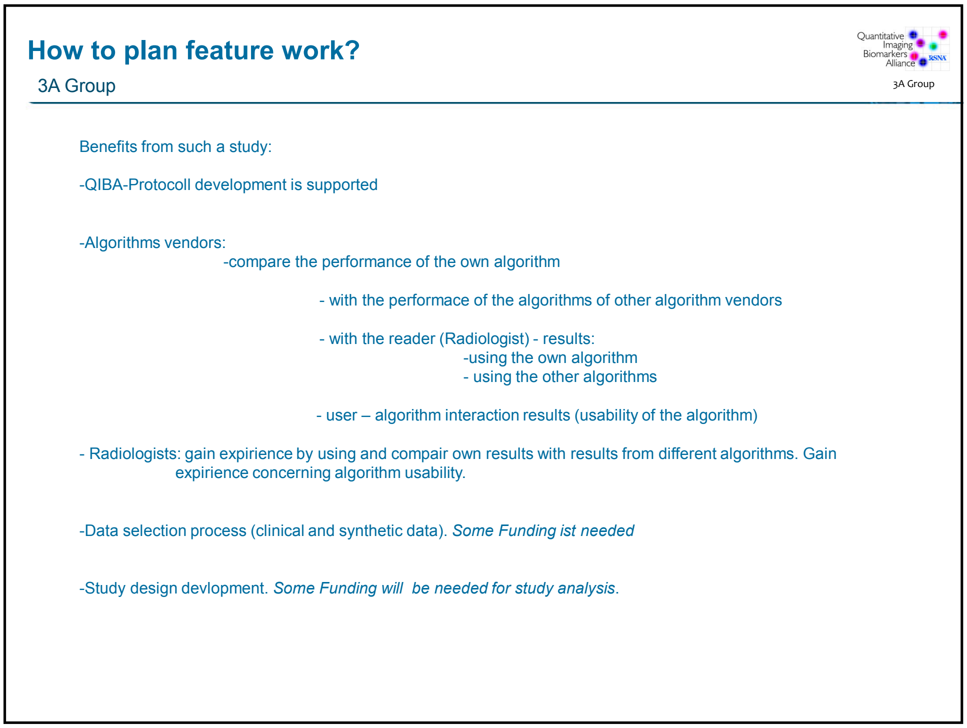
Study Components:

-Softwareplatformdevelopment (Qi-Bench?, any other possibility?) *Some Funding ist is probably needed*

-Algorithms (asking the vendors, if they want to apply theis software for this kind of study?)

-Data selection process, synthetic data is included. *Some Funding ist needed*

-Study design devlopment. *Some Funding will be needed for study analysis*.



How to plan feature work?

3A Group

Benefits from such a study:

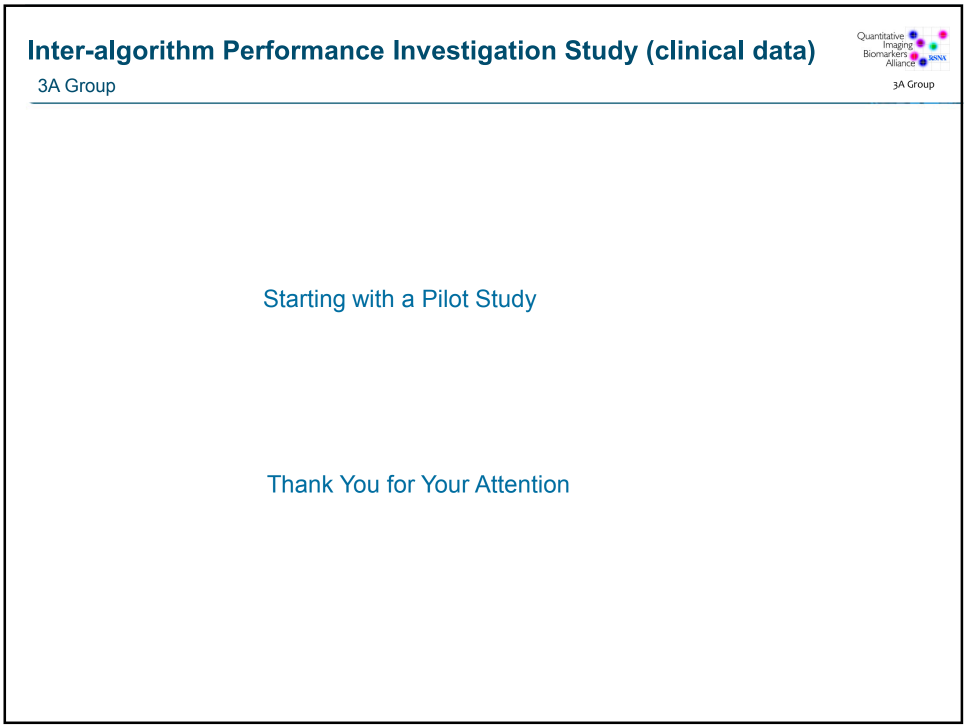
-QIBA-Protocoll development is supported

-Algorithms vendors:
-compare the performance of the own algorithm
- with the performace of the algorithms of other algorithm vendors
- with the reader (Radiologist) - results:
-using the own algorithm
- using the other algorithms
- user – algorithm interaction results (usability of the algorithm)

- Radiologists: gain expirience by using and compair own results with results from different algorithms. Gain expirience concerning algorithm usability.

-Data selection process (clinical and synthetic data). *Some Funding ist needed*

-Study design devlopment. *Some Funding will be needed for study analysis*.



Inter-algorithm Performance Investigation Study (clinical data)

3A Group

Starting with a Pilot Study

Thank You for Your Attention