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

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

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

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

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

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

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

Acad Radiol 2015; 22:1393–1408

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

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

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

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

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

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

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

MATERIALS AND METHODS

Data collection

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

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

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

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