Data sets for the qualification of volumetric CT as a quantitative imaging biomarker in lung cancer◊

A. J. Buckler,[1,*] L. H. Schwartz,[2] N. Petrick,[3] M. McNitt-Gray,[4] B. Zhao,[2] C. Fenimore,[5] A. P. Reeves,[6] P. D. Mozley,[7,8] and R. S. Avila[9]

[1]*Buckler Biomedical LLC, Wenham, MA, USA*
[2]*Columbia University, New York, NY, USA*
[3]*Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD, USA*
[4]*David Geffen School of Medicine at UCLA, Los Angeles, CA, USA*
[5]*National Institute of Standards and Technology, Gaithersburg, MD, USA*
[6]*Cornell University, Ithaca, NY, USA*
[7]*Merck Research Laboratories, West Point, PA, USA*
[8]*Extended Pharma Imaging Group*
[9]*Kitware Inc., Clifton Park, NY, USA*
[*]andrew@bucklerbiomedical.com

Abstract: The drug development industry is faced with increasing costs and decreasing success rates. New ways to understand biology as well as the increasing interest in personalized treatments for smaller patient segments requires new capabilities for the rapid assessment of treatment responses. Deployment of qualified imaging biomarkers lags apparent technology capabilities. The lack of consensus methods and qualification evidence needed for large-scale multi-center trials, as well as the standardization that allows them, are widely acknowledged to be the limiting factors. The current fragmentation in imaging vendor offerings, coupled with the independent activities of individual biopharmaceutical companies and their contract research organizations (CROs), may stand in the way of the greater opportunity were these efforts to be drawn together. A preliminary report, "Volumetric CT: a potential biomarker of response," of the Quantitative Imaging Biomarkers Alliance (QIBA) activity was presented at the Medical Imaging Continuum: Path Forward for Advancing the Uses of Medical Imaging in the Development of New Biopharmaceutical Products meeting of the Extended Pharmaceutical Research and Manufacturers of America (PhRMA) Imaging Group sponsored by the Drug Information Agency (DIA) in October 2008. The clinical context in Lung Cancer and a methodology for approaching the qualification of volumetric CT as a biomarker has since been reported [Acad. Radiol. 17, 100–106, 107–115 (2010)]. This report reviews the effort to collect and utilize publicly available data sets to provide a transparent environment in which to pursue the qualification activities in such a way as to allow independent peer review and verification of results. This article focuses specifically on our role as stewards of image sets for developing new tools.

© 2010 Optical Society of America

OCIS codes: (110.2960) image analysis; (100.3008) image recognition, algorithms and filters

◊Data sets associated with this article are available at http://hdl.handle.net/10376/1523. **Links such as "View 1" that appear in figure captions and elsewhere will launch custom** data views if ISP software is present.

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

References and links

1. E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)," Eur. J. Cancer 45(2), 228–247 (2009).
2. N. R. Bogot, E. A. Kazerooni, A. M. Kelly, L. E. Quint, B. Desjardins, and B. Nan, "Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods," Acad. Radiol. 12(8), 948–956 (2005).
3. P. D. Mozley, L. H. Schwartz, C. Bendtsen, B. Zhao, N. Petrick, A. J. Buckler, "Change in lung tumor volume as a biomarker of treatment response: A critical review of the evidence," Ann. Oncol. (to be published).
4. K. Marten, F. Auer, S. Schmidt, G. Kohl, E. J. Rummeny, and C. Engelke, "Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria," Eur. Radiol. 16(4), 781–790 (2006).
5. J. J. Erasmus, G. W. Gladish, L. Broemeling, B. S. Sabloff, M. T. Truong, R. S. Herbst, and R. F. Munden, "Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response," J. Clin. Oncol. 21(13), 2574–2582 (2003).
6. J. E. Munzenrider, M. Pilepich, J. B. Rene-Ferrero, I. Tchakarova, and B. L. Carter, "Use of body scanner in radiotherapy treatment planning," Cancer 40(1), 170–179 (1977).
7. J. M. Quivey, J. R. Castro, G. T. Chen, A. Moss, and W. M. Marks, "Computerized tomography in the quantitative assessment of tumour response," Br. J. Cancer Suppl. 4, 30–34 (1980).
8. C. G. Moertel, and J. A. Hanley, "The effect of measuring error on the results of therapeutic trials in advanced cancer," Cancer 38(1), 388–394 (1976).
9. M.-P. Revel, C. Lefort, A. Bissery, M. Bienvenu, L. Aycard, G. Chatellier, and G. Frija, "Pulmonary nodules: preliminary experience with three-dimensional evaluation," Radiology 231(2), 459–466 (2004).
10. B. Zhao, L. H. Schwartz, C. S. Moskowitz, M. S. Ginsberg, N. A. Rizvi, and M. G. Kris, "Lung cancer: computerized quantification of tumor response--initial results," Radiology 241(3), 892–898 (2006).
11. B. Zhao, G. R. Oxnard, P. Guo, *et al.*, "A pilot study comparing computerized volume measurement with diameter measurement as an early biomarker of the biologic activity of EGFR targeted therapy," IASLC 13th World Conference on Lung Cancer, July 31–August 4, 2009, San Francisco, California.
12. L. Schwartz, S. Curran, R. Trocola, J. Randazzo, D. Ilson, D. Kelsen, and M. Shah, "Volumetric 3D CT analysis–an early predictor of response to therapy," J. Clin. Oncol. 25(18S), 4576 (2007).
13. N. Altorki, J. Heymach, M. Guarino, P. Lee, E. Felip, T. Bauer, S. Swann, D. Roychowdhury, L. H. Ottesen, and D. Yankelevitz, "Phase II study of pazopanib (GW786034) given preoperatively in stage I–II non-small cell lung cancer (NSCLC): a proof-of-concept study," Ann. Oncol. 19(Supplement 8), 124 (2008).
14. L. P. Clarke, R. D. Sriram, and L. B. Schilling, "Imaging as a biomarker: standards for change measurements in therapy workshop summary," Acad. Radiol. 15(4), 501–530 (2008).
15. G. McLennan, L. P. Clarke, and R. J. Hohl, "Imaging as a biomarker for therapy response: cancer as a prototype for the creation of research resources," Clin. Pharmacol. Ther. 84(4), 433–436 (2008).
16. S. G. Armato 3rd, C. R. Meyer, M. F. Mcnitt-Gray, G. McLennan, A. P. Reeves, B. Y. Croft, L. P. Clarke; RIDER Research Group, "The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: a resource for the development of change-analysis software," Clin. Pharmacol. Ther. 84(4), 448–456 (2008).
17. N. Petrick, D. G. Brown, O. Suleiman, and K. J. Myers, "Imaging as a tumor biomarker in oncology drug trials for lung cancer: the FDA perspective," Clin. Pharmacol. Ther. 84(4), 523–525 (2008).
18. M. M. Goodsitt, H.-P. Chan, T. W. Way, S. C. Larson, E. G. Christodoulou, and J. Kim, "Accuracy of the CT numbers of simulated lung nodules imaged with multi-detector CT scanners," Med. Phys. 33(8), 3006–3017 (2006).
19. M. A. Gavrielides, L. M. Kinnard, K. J. Myers, J. Peregoy, W. F. Pritchard, R. Zeng, J. Esparza, J. Karanian, and N. Petrick, "A resource for the development of methodologies for lung nodule size estimation: database of thoracic CT scans of an anthropomorphic phantom," Opt. Express 18(14), 15244–15255 (2010).
20. E. Nioutsikou, N. Richard, J. Symonds-Tayler, J. L. Bedford, and S. Webb, "Quantifying the effect of respiratory motion on lung tumour dosimetry with the aid of a breathing phantom with deforming lungs," Phys. Med. Biol. 51, 3359–3374 (2006).
21. T. W. Way, H.-P. Chan, M. M. Goodsitt, B. Sahiner, L. M. Hadjiiski, C. Zhou, and A. Chughtai, "Effect of CT scanning parameters on volumetric measurements of pulmonary nodules by 3D active contour segmentation: a phantom study," Phys. Med. Biol. 53(5), 1295–1312 (2008).
22. M. A. Gavrielides, L. M. Kinnard, K. J. Myers, and N. Petrick, "Noncalcified lung nodules: volumetric assessment with thoracic CT," Radiology 251(1), 26–37 (2009).
23. M. A. Gavrielides, R. Zeng, L. M. Kinnard, K. J. Myers, and N. Petrick, "A template-based approach for the analysis of lung nodules in a volumetric CT phantom study," Proc. SPIE 7260, 726009–726011 (2009).
24. M. F. McNitt-Gray, L. M. Bidaut, S. G. Armato, C. R. Meyer, M. A. Gavrielides, C. Fenimore, G. McLennan, N. Petrick, B. Zhao, A. P. Reeves, R. Beichel, H. J. Kim, and L. Kinnard, "Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error," Transl. Oncol. 2(4), 216–222 (2009).

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

25. C. R. Meyer, S. G. Armato, C. P. Fenimore, G. McLennan, L. M. Bidaut, D. P. Barboriak, M. A. Gavrielides, E. F. Jackson, M. F. McNitt-Gray, P. E. Kinahan, N. Petrick, and B. Zhao, "Quantitative imaging to assess tumor response to therapy: common themes of measurement, truth data, and error sources," Transl. Oncol. 2(4), 198–210 (2009).
26. B. Zhao, L. P. James, C. S. Moskowitz, P. Guo, M. S. Ginsberg, R. A. Lefkowitz, Y. Qin, G. J. Riely, M. G. Kris, and L. H. Schwartz, "Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer," Radiology 252(1), 263–272 (2009).
27. A. P. Reeves, A. M. Biancardi, D. Yankelevitz, S. Fotin, B. M. Keller, A. Jirapatnakul, and J. Lee, "A public image database to support research in computer aided diagnosis," in *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sept. 2009, pp. 3715–3718.
28. http://www.via.cornell.edu/databases/crpf.html.
29. http://www.grand-challenge.org/index.php/MICCAI_2010_Workshop.
30. https://wiki.nci.nih.gov/display/Imaging/Algorithm+Validation+Toolkit+(AVT).
31. http://www.itl.nist.gov/iad/894.05/biochange2008/Biochange2008-webpage.htm.
32. A. P. Reeves, A. C. Jirapatnakul, A. M. Biancardi, *et al.*, "The VOLCANO'09 challenge: preliminary results," in *Second International Workshop of Pulmonary Image Analysis*, Sept. 2009, pp. 353–364.
33. http://preventcancer.org/.
34. A. P. Reeves, A. M. Biancardi, T. V. Apanasovich, C. R. Meyer, H. MacMahon, E. J. van Beek, E. A. Kazerooni, D. Yankelevitz, M. F. McNitt-Gray, G. McLennan, S. G. Armato 3rd, C. I. Henschke, D. R. Aberle, B. Y. Croft, and L. P. Clarke, "The Lung Image Database Consortium (LIDC): a comparison of different size metrics for pulmonary nodule measurements," Acad. Radiol. 14(12), 1475–1485 (2007).
35. http://www.via.cornell.edu/volcaman/ Draft version of the VOLCAMAN study.
36. http://qibawiki.rsna.org/index.php?title=Volumetric_CT.

1. Unmet Medical Needs as Business Drivers for Qualifying Quantitative Imaging

X-ray computed tomography (CT) is a three dimensional imaging technique that can non-invasively portray internal anatomy and pathological masses. Subjective impressions of changes in tumor masses based on serially acquired CT scans can be sufficient for making sound judgments about the effects of treatment when therapy is so robustly beneficial that improvements are conspicuous, or when the therapy fails so completely that disease progression is obvious. However, as the "war on cancer" matures from hopes of curing some of these diseases into aspirations of managing morbidity over progressively longer and longer time horizons, needs for rapidly assessing small changes in tumors and quantifying the incremental value of new drugs are becoming increasingly important. Problems with qualitative "reads" that emerge when treatment effects are small or measured over short time intervals include inadequate levels of inter-reader concordance. Discordance among "readers" has led to skepticism about medical imaging as a reliable biomarker of response, as well as confusion about whether some investigational new drugs should be advanced in development settings or approved by regulatory authorities for general use in practice settings.

For an individual patient in an ordinary medical setting, being prescribed a marketed treatment regimen that has been established as sufficiently safe and effective in large populations is analogous to starting a personal clinical trial. This is because even the best treatment regimens fail in a some portion of patients with the disease, and even relatively safe therapies cause serious side effects in some people. These principles seem to hold for all treatments, and particularly for anti-neoplastic therapies. Patients want to know as soon as possible if their new-to-them treatment is conveying benefits. If it is not, then they want to launch a search for alternatives as soon as possible.

No one wants to waste time, effort, and money on treatments that are not helpful. From this perspective, the interests of individual patients and third party payers seem highly concordant. Many new treatments are expensive. Some are cost effective in individuals, but less so in large populations. New methods are needed to determine who is who. Until definitive enrichment tools are developed for matching individual patients to specific treatments, the early assessment of response will remain the primary mechanism for sparing resources.

Biopharmaceutical enterprises view clinical trials of novel products the same way as the other stakeholders in the management of cancer. Like individual patients, industry wants its

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

products to succeed for the patients who use them, and as a consequence, produce a net-positive return on investment. More sensitive biomarkers of response would allow industry to reduce the number of patients required to test new products, as well as decrease the amount of time that patients need to remain on-study. The net effect would increase the number of new treatments for unmet medical needs that reach the market and make a positive impact on human health, primarily by allowing investigational new treatments to fail faster than is currently possible in clinical trials that use survival or clinical signs of progression as their endpoints.

Response Evaluation Criteria in Solid Tumors (RECIST) [1] is a quantitative image analysis technique. It is currently based on using the longest, in-plane diameter of a tumor as a proxy for its mass. Changes in longest diameters (LDs) during the course of illness usually reflect changes in health status, as decreases should correspond to remission, and increases should reflect progression of the disease. There are many reports of using LD-based RECIST to successfully distinguish between different treatment arms in clinical trials [2]. However, concerns have been raised about relying on measurements of LDs on only one axial slice per tumor [3]. Problems with the precision of measurement have been described [2,4,5] As a consequence of measurement variability, the categorical response of Stable Disease is broad. Decreases in LDs of 30% or more are required for changing an assessment category from Stable Disease to partial response, while increases in LDs of 20% or more are required for triggering assessments of progressive disease. For tumors that can be modeled as spheres, these changes correspond to changes of about −66% and + 73%, respectively. Because these thresholds are relatively large and can take a long time for some patients to cross them, there is a need for more sensitive methods for making assessments of response with confidence.

The point is illustrated in Fig. 1, which shows the actual data for a patient who participated in a clinical trial of a new treatment for advanced stage lung cancer. Rules for making assessments based on changes in the longest diameters of the target lesions require the clinical course of illness to be classified as one of prolonged Stable Disease. As a consequence, the subject added little analytical power needed to distinguish between the two arms of the trial. In retrospect, volumetric image analysis suggests that this patient had an initial response to treatment, but could have come off trial and switched to a new treatment several months before changes in unidimensional line-lengths met criteria for Progressive Disease.

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

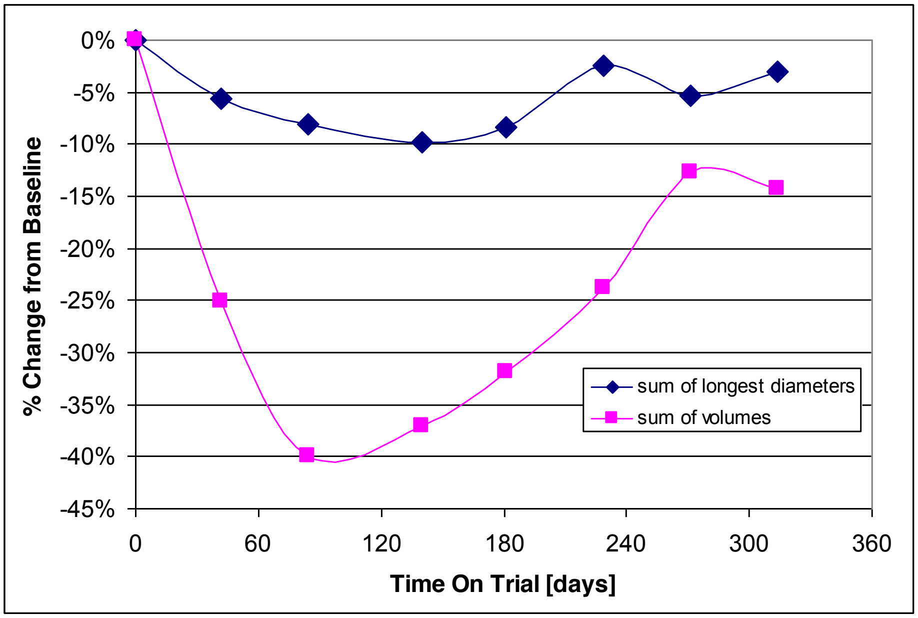


Fig. 1. Data for a patient who participated in a clinical trial of a new treatment for advanced stage lung cancer

All of the stakeholders lose when therapeutic benefits are under appreciated, or there are delays in diagnosing Progressive Disease. The hypothesis that quantifying whole tumor volumes as the basis for response evaluation criteria is actually quite old [6,7], and in fact preceded the advent of CT [8]. The question about replacing LDs with volumes is being re-posed in this work because it might be that improvements in image quality and image analysis now make it technically feasible to quantify some tumor volumes with continuously improving precision and accuracy. The need to test the hypothesis again seems urgent in part because a number of investigators have shown that the measurements of whole tumor volumes can be more precise [9] and sensitive [10–13] than the measurement of the corresponding LDs.

2. Methods

It is widely recognized that significant advances in imaging technology have led to an increasingly important role for imaging in diagnosis, staging, guiding systemic, local, or interventional therapies, and monitoring responses to treatment. However, development of imaging technologies is expensive, and early phase justification of effectiveness, before commercial viability is established, can be difficult. There is an emerging consensus that a cooperative atmosphere must be developed among the biopharmaceutical industry, the imaging device manufacturers, government funding agencies, and regulatory authorities, as well as scientists in a wide range of fields, to cost effectively select and qualify mature quantitative imaging methods as biomarkers for the measurement of response to therapy.

The development of public resources and open source tools for imaging as a biomarker using X-ray CT was re-invigorated by the National Cancer Institute (NCI), National Institute of Biomedical Engineering and Bioengineering (NIBIB), Food and Drug Administration (FDA) and National Institute of Standards and Technology (NIST) in 2005, which included collaboration with the Radiological Society of North America (RSNA) [12,14–17]. This earlier work prompted the organization of an inter-federal agency workshop held at NIST in September 2006, which addressed physical standards for imaging as a biomarker [3]. Stakeholders from academia, industry, and scientific imaging societies including RSNA,

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010
(C) 2010 OSA

American Association of Physicists in Medicine (AAPM), Society of Nuclear Medicine (SNM), and the International Society for Magnetic Resonance in Medicine (ISMRM) proposed a model similar to the "Integrating the Healthcare Enterprise" (IHE) paradigm to engage industry stakeholders in this research area.

At its annual meeting in 2007, RSNA created the Quantitative Imaging Biomarker Alliance (QIBA) to investigate the role of quantitative imaging methods in CT, MRI and PET as potential biomarkers in evaluating disease and responses to treatment. The alliance has formed technical committees of representatives from the instrumentation manufacturers, software developers, imaging professionals in the pharmaceutical industry, radiologists from the imaging contract research organizations (CROs), officers in regulatory agencies, governmental research organizations, imaging scientists, and professional imaging society representatives. One of the technical committees is referred to as the "Quantitative CT Technical Committee."

The Quantitative CT Technical Committee is engaged to produce alternative methods of response assessment, based on volumetric image acquisition and analysis, which will be accepted through appropriate regulatory pathways as predictors of clinical benefits, such as overall survival (OS). The first specific aim compares time-dependent outcome measures based on uni-dimensional longest diameters to analogous endpoints based on 3D volumetric image analyses. The expectation is that these alternative methods would be adopted if they require fewer enrollees in clinical trials, shorten time on trial for each subject who will ultimately fail to benefit from treatment, decrease the length of time required to conduct trials, and/or provide better correlations with actual clinical outcomes.

The Committee was formed to include practicing clinicians, professional society leaders, regulatory officers, pharmaceutical industry representatives, imaging scientists, and imaging device industry representatives. The principal value of the effort is to help converge the interests and effort of many stakeholders.

Long-Term Goals are to establish processes and profiles that will eventually lead to the acceptance by the imaging community, clinical trial industry, and regulatory agencies, of 3D volumetric CT as a surrogate end-point for changes in the health status of patients.

Specific Aims are to develop the capability to meet targeted levels of accuracy and reproducibility for the quantification of anatomical structures, such as neoplastic masses. This in turn requires identifying and creating mitigation strategies for all significant sources of variability in these measurements as necessary to meet the targets.

Context is that this work is being conducted under the aegis of the RSNA's QIBA in collaboration with FDA's Division of Applied Math/ Office of Science and Engineering Laboratories (OSEL)/ Center for Devices and Radiological Health (CDRH), NCI, NIST, American College of Radiology Imaging Network (ACRIN), major imaging equipment manufacturers (Philips, GE, Siemens, Toshiba, etc.), the Extended Pharmaceutical Research and Manufacturers of America (PhRMA) Imaging Group, and others.

Constraint is that this work depends on the collaboration of, and must demonstrate benefit to, the imaging industry, the pharma industry, the academic research community, individuals with cancer, and the clinical community. The benefits must be robust to justify the increased time and effort required when compared to qualitative impressions, as well as satisfy the requirements of the regulatory agencies. Our approach is to converge scientific analysis in a way that encourages vendor participation while meeting current biopharmaceutical industry needs.

Our ultimate goal is the use of these biomarkers on typical imaging systems in the practice of medicine.

3. Results to Date

The QIBA initiative has explored a number of issues and opportunities to improve research and development of volumetric CT therapy assessment methods. To accomplish this, it has

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

been essential to obtain and analyze a wide range of image data collections that span clinical concepts and challenges, fundamentals of image acquisition, and opportunities to better perform the evaluation of algorithm performance. The sections that follow describe these data collections and the important insights each collection provides to the research community.

3.1 Understanding Performance on Phantoms

One approach to efficiently develop and evaluate the applicability of a quantitative imaging biomarker is to investigate the biomarker's performance with phantom data. Phantom image data can come in many forms including imaging simple lesion-like objects on flat backgrounds or imaging anthropomorphic phantoms containing realistic structure, complex synthetic lesions, and realistic physiology. Figure 2 shows three different examples of lung and chest phantoms from the literature, including a tissue equivalent tissue equivalent thorax section phantom (Fig. 2a), an anthropomorphic chest phantom (Fig. 2b, and a mechanical breathing phantom (Fig. 2c) [18–20]

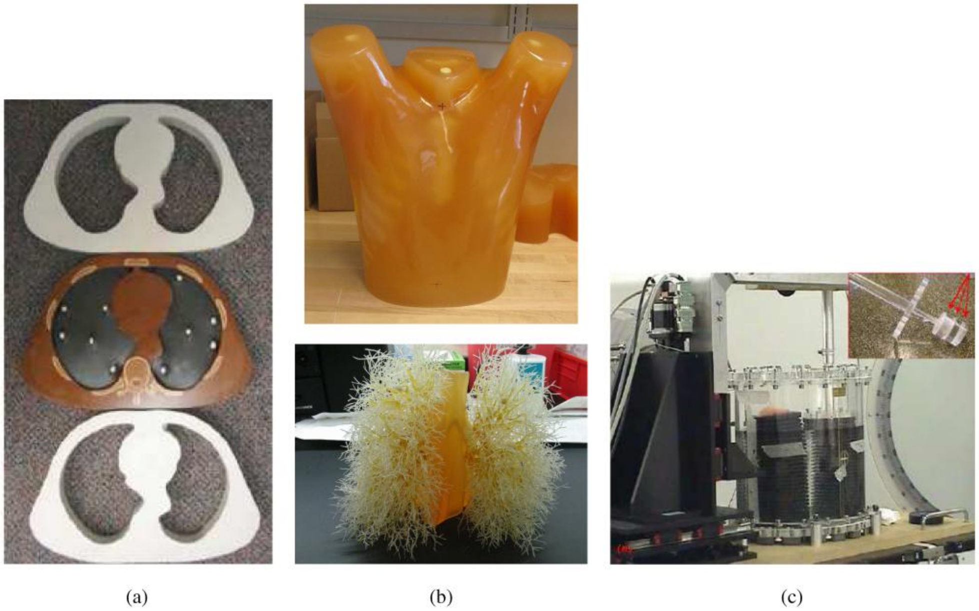


Fig. 2. (a) tissue equivalent thorax section phantom (center) containing 9.5 mm diameter simulated spherical lung nodules, with two water-equivalent bolus sections (top and bottom), (b) the exterior shell of an anthropomorphic thoracic phantom and its vasculature insert; and (c) a mechanical lung phantom used to simulate breathing. Images in (a)-(c) are reprinted with permission.

Although phantoms are different from real patients in many ways, phantom studies allow for a systematic analysis of biomarker performance against a known reference standard and under a range of imaging conditions. This type of systematic analysis would be virtually impossible to conduct using patient scans because of dose concerns, variability in patients, motion artifacts, and lack of a definitive truth standard [21]. While phantom studies are unlikely to serve as a complete replacement for evaluating a new biomarker on patient data, they may serve at least three important functions. One is to quickly triage potential imaging biomarkers, so that time is not wasted evaluating biomarkers that have little potential for providing reliable quantitative measurements. New biomarkers that don't perform well with idealized phantom data are unlikely to perform well in patients whose diseases are well modeled by the phantom. For those imaging biomarkers that do show promise, a second function of phantom data could be to systematically probe how biomarker performance is impacted by variations in imaging hardware and image acquisition protocols. Again, this type of systematic evaluation of a biomarker is virtually impossible to conduct with patient data, even within a clinical trial, because of the large variability in manifestations of disease both within and among patients. Finally, a third contribution of phantom studies could be in the design of clinical trials incorporating an imaging biomarker. By first understanding how variations in image acquisition affect the reliability of the quantitative measurement through phantom studies [22], it becomes possible to develop appropriate imaging standards as well as determining a minimum number of patients required to overcome the variability implicit when implementing the imaging biomarker. Additional patients, above this minimum level, would be necessary to overcome patient variability as well as other sources of error in any particular trial.

Gavrielides et al. describes CT image data for an anthropomorphic thorax phantom containing synthetic lung nodules [22]. These data were collected by the U.S. Food and Drug Administration (FDA) to evaluate various lesions size measurement algorithms, and to develop a more complete understanding of how algorithm performance changes with variations in CT acquisition protocols and imaging hardware. Figure 2(b) shows the thorax phantom and vasculature lung inserts to which synthetic nodules were attached and then imaged within the data set. The phantom was scanned with a Philips 16-row scanner (Mx8000 IDT, Philips Healthcare, Andover, MA) and a Siemens 64-row scanner (Somatom 64, Siemens Medical Solutions USA, Inc., Malvern, PA). The data were collected using a factorial design so that a large number of combinations of exposure, pitch, slice collimation, reconstruction kernels and slice thickness were collected for both simple spherical nodules ad well as more complex ovoid, lobulated and spiculated synthetic nodules. Figure 3 shows a complete CT scan of the phantom with seven spherical nodules of various sizes and densities attached to the vasculature insert.

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

(C) 2010 OSA

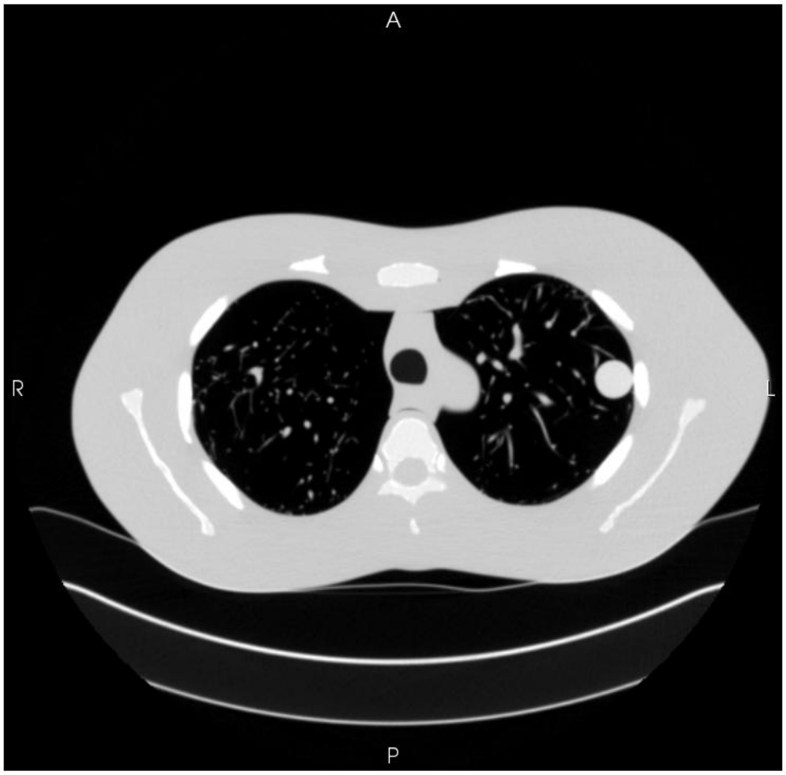


Fig. 3. CT scan from acquisition 9111 of the FDA phantom data set. The thorax phantom contained six spherical nodules (20 mm diameter with −630 HU density; 5 mm. 8 mm, 10 mm, 20 mm and 40 mm diameter with −10 HU density; 10 mm and 20 mm with + 100 HU density). The scan was acquired on a Philips Mx8000 IDT scan at 120 KVp and 200 mAs using a 16x0.75 collimation. 1.5 mm reconstruction thickness, 0.75 reconstruction increment, pitch of 1.2 and a medium reconstruction kernel (View 1).

The FDA thorax phantom CT data described in [22] can be used as a resource for the development and assessment of lung nodule sizing algorithms. Both the bias and variance associated with a nodule sizing method can be obtained because the reference standard for nodule size as well as repeat exposures are included as part of the data set. This makes the data ideal for comparing various size estimation algorithms. The data are also useful for developing new size estimation methods [23] as well as developing appropriate assessment methodologies for comparing algorithms. These as well as various other applications of the phantom data are discussed in more detail in [22].

Evaluation of imaging biomarkers with phantom data is one important component in the qualification of these biomarkers in both drug trials and clinical practice. Clearly, phantom data have limitations because they do not match the diversity or complexity of real patients. This strongly suggests that testing on patient data will be necessary at some point in the development process, but also that phantom data can be a very effective tool in both streamlining the development process and maximizing the utility of patient image data.

3.2 Clinical Data Resources

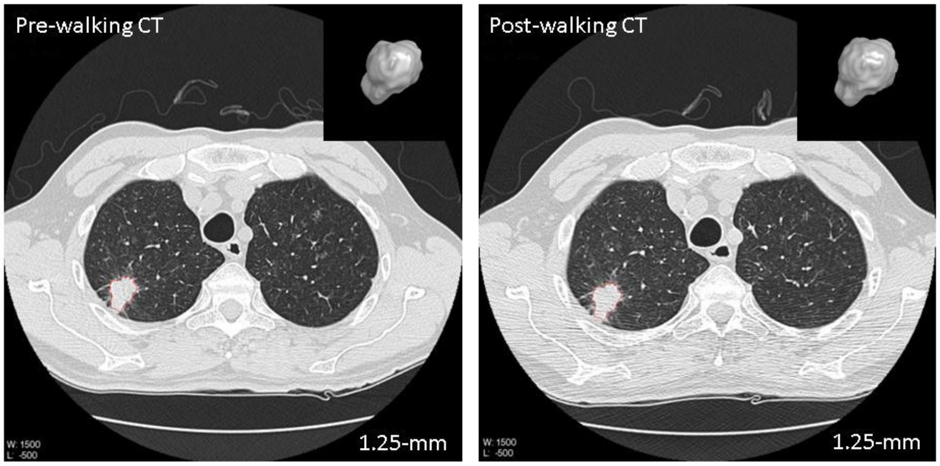
There have been considerable efforts to create publicly available sets of image data to assist in some of the efforts related to quantitative imaging of disease. These data sets represent an important aspect in establishing quantitative imaging methods as they serve as reference data sets against which investigators and researchers may be able to benchmark and compare their

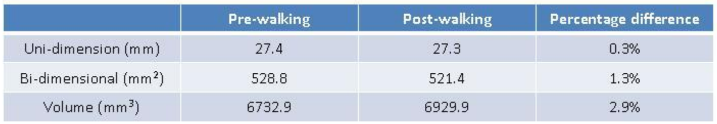
#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

measurement algorithms. Several data sets are now available, primarily through the NCI-funded Reference Image Database to Evaluate Response to Therapy (RIDER) [6,24,25]

3.2.1 Same-day repeat CT study in NSCLC patients

The first data set to describe is the No-Change data set provided by Memorial Sloan Kettering Cancer center [26]. In this study, 32 patients with Non-Small Cell Lung Cancer (NSCLC) were consented and scanned twice within 15 minutes on the same scanner with the same imaging acquisition protocol. An example of the scans is shown in Fig. 4. The scanners were either LightSpeed 16 or VCT 64 (GE Healthcare, Milwaukee, WI). Images of each scan were reconstructed at 1.25mm slice interval without overlap. This experiment represents repeat scans under a presumed "no change" condition. Tumor differences measured between the two scans can be considered as measurement variation/error that is possibly caused by intrinsic variance in the CT scanning device, errors in the image processing system, differences in patient positioning, patient inspiration level, etc. Because this data set does contain the same lesions acquired on two repeat CT scans under identical parameter settings in a short time period, it can be used to investigate minimum detectable changes on the state-of-the-art CT scanners by using advanced measurement tools, the information needed to define tumor response and progression. These data sets have been made publicly available through the NBIA web archive (http://ncia.nci.nih.gov/) and can specifically be accessed through the shared list identified (exactly) as "MSK_coffee_break_CT," which will contain the 64 series.





	Pre-walking	Post-walking	Percentage difference
Uni-dimension (mm)	27.4	27.3	0.3%
Bi-dimensional (mm²)	528.8	521.4	1.3%
Volume (mm³)	6732.9	6929.9	2.9%

Courtesy of Laboratory for Computational Image Analysis, Columbia University Medical Center

Fig. 4. An example taken from the same-day repeat CT study. Computer-aided tumor measurements were different on the two repeat CT scans even if there was no biological change of the tumor (View 2).

3.2.2 CT lung studies at different time intervals

In another RIDER project related study, serial CT scan images of patients with known tumors in the lungs (both primary and metastatic lesions) were submitted to NBIA under the RIDER

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

collection (Fig. 5). Each case had at least 2 image data sets from different time points; many had 3 or more time points. These cases were collected from UT-MD Anderson Cancer Center and Memorial Sloan-Kettering Cancer Center, as part of their clinical operation. There was no specific attempt to tightly control the imaging parameters between studies for these patients.

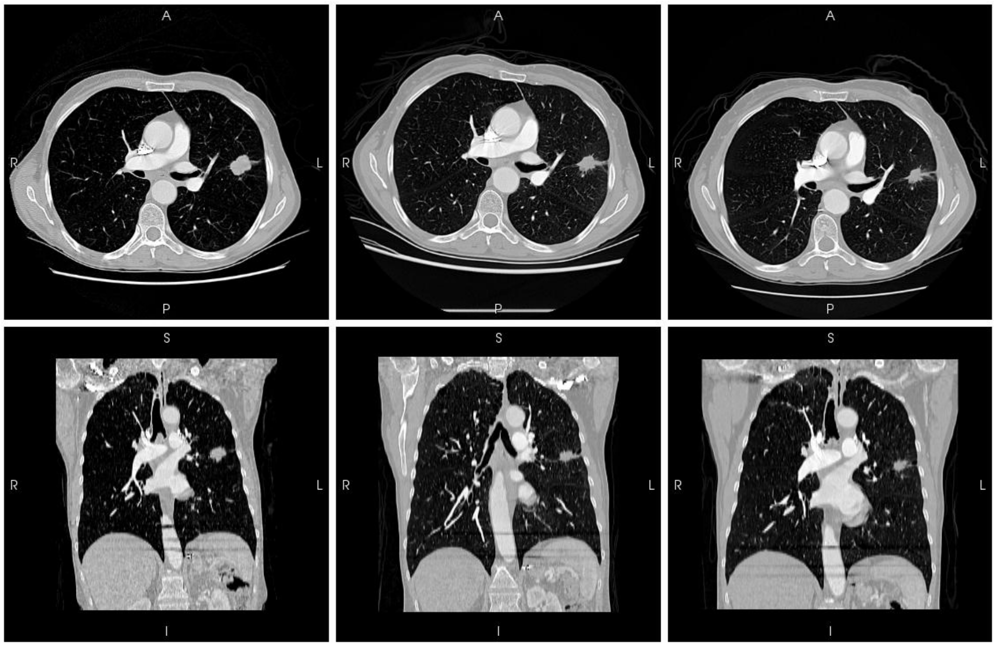


Fig. 5. Longitudinal Scans where Patient has Known Tumor (View 3).

Another public resource for clinical CT image data is the Public Lung Database to Address Drug Response [27,28] This data set contains a number of different exemplar CT image sets including cases with at least two scans having manual volumetric boundary markings and cases with at least two scans recorded in the same session (zero-change) as part of a biopsy procedure that are documented with a semi-automated lesion measuring algorithm. These cases were collected from the Weill Cornell Medical College as part of their clinical operation.

While these reference data sets cannot be used to quantify the accuracy of measurement, they are a tremendous resource for researchers who need to characterize the precision of new quantitative imaging methods. They can be used to investigate the minimum detectable change (using the cases with no change) as well as different sources of variance (both sets).

3.3 Algorithm Evaluation Systems

We expect that computer assisted methods for measurement will aid the physician with respect to accuracy and precision of lesion measurements. One principal goal in evaluating such methods is to support the improvement of algorithms by providing developers a resource for identifying the strengths and weaknesses of their methods. Similar evaluations have been applied to computer vision methods for biometric-based identification, such as face and gait recognition, as well as in medical imaging. We build on the accomplishments of other efforts, such as Medical Image Computing and Computer-Assisted Intervention (MICCAI)'s algorithm challenges [29], the National Cancer Institute (NCI) Cancer Bioinformatics Grid (caBIG)'s Algorithm Validation Toolkit (AVT) project [30], algorithm evaluation for commercial detection (rather than measurement) products (e.g., mammo, lung and colon cancer), and measurement in other quantitative medical fields (such as functional MRI for neuroimaging).

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

For the clinical use of the volumetric image biomarker the most relevant measurement is the relative change in lesion size over some time interval. As has been stated before, it is critical to know when a measured change in size is statistically significantly greater than the measurement error (i.e., represents an actual change in the lesion); secondly we would like to know the precision of the size change measurement. To explore these issues in the context of computer algorithms and real lesions rather than phantoms, studies have been conducted on selected data sets of pairs of lesions to evaluate how different computer algorithms compare on a standardized data set.

An evaluation of this type was Biochange'08, which invited medical software developers to apply their stand-alone software or computer-assisted markup tools to measure the change in pulmonary lesions. The lung CT data was drawn from the RIDER database of patients with known lung tumors, described the above section *CT lung studies at different time intervals* and from the CT imagery of the FDA's anthropomorphic phantom described earlier [31]. This pilot study provided algorithm and software developers with 13 cases, each having series at 2 time points. Seven cases were clinical, all with 5.0 mm slice thickness and acquired at intervals of weeks to months. The clinical cases were chosen from among 23 RIDER cases for which markup by 2 radiologists of lesion diameters is available (NBIA). There were six phantom nodule pairs from studies of the FDA phantom, having slice thicknesses of 3.0 mm and 0.8 mm.

Biochange'08 was designed as a pilot, a proof of concept for the evaluation process. For each lesion, participants were provided with a seed point in a region-of-interest. Three organizations participated and provided 4 sets of change measurements. Three of the submissions involved semi-automated segmentation tools, while one was stand-alone software without user interaction. The study required the participant to submit a measure of change for each case. While this permitted the use of any change metric, for example ones based on one- or two-dimensional measurement, each participant submitted the fractional change for volume and provided volume measurements at both time points.

In the analysis, the markup was used as a reference against which the submitted results were compared. The limited size of the study did not support statistically significant findings about the differences between the submissions but did suggest some tentative conclusions regarding the comparison of diameter measurement markup of axial slices and computer assisted change measurement. The phantom data provided insight into the effects of slice thickness on the measurement of volume change.

The data suggest the various software submissions achieve agreement comparable to that achieved between the two radiologists. As can seen in Fig. 6, the two groups reached similar conclusions regarding categorical change in the lesions. In particular, there were 6 cases for which the two radiologists agreed on the categorical assessment of change (response/stable disease/disease progression) based on the diameter measurements on axial slices. The readers disagreed in one case, RIDER 2 as seen in Fig. 6. Using categorical 3-dimensional thresholds derived from the diameter measurements on axial slices criteria, the 4 submissions obtained results similar to those of the radiologists: agreeing with each other in 5 of the 6 cases, while disagreeing in one, RIDER 6 in Fig. 6. The two cases of disagreement occurred on lesions involved, in one case, with the mediastinum and in the other, with the lung wall at the apex.

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

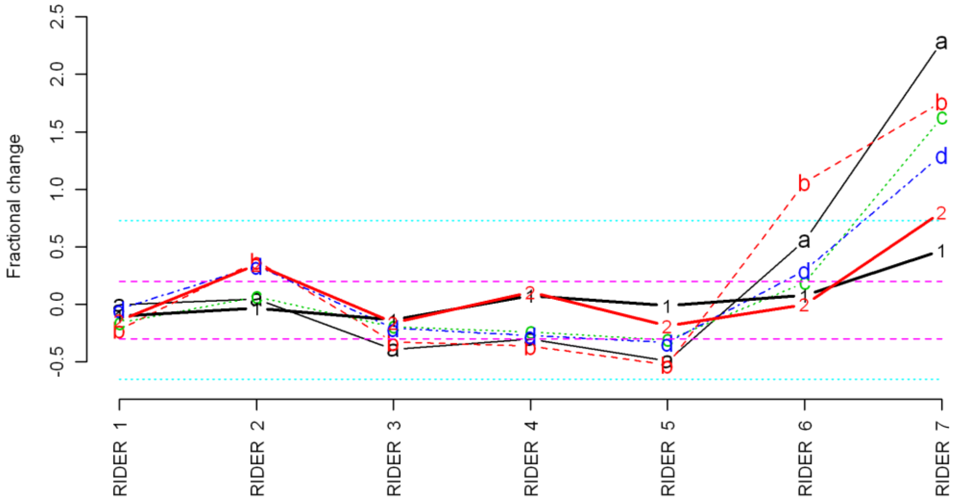


Fig. 6. Comparison of (non-dimensional) reported fractional volume change of RIDER image pairs for four Biochange'08 participants (a, b, c, and d) to the computed fractional change (in diameter) based on the two readers' diameter readings (1 and 2). The two dashed (purple) lines correspond to the RECIST criteria for progressive disease (20% increase in diameter, top), and partial response (30% decrease in diameter). The readers disagree on one case, RIDER 2. The two dotted (light blue) lines correspond to 73% increase in volume (top) and 65% decrease in volume. The software submissions also disagree on a single case, RIDER 6 (View 4).

The phantom nodules were scanned in both thin- and thick-slice series (0.8 and 3.0 mm). The phantom nodule comparisons were between two scans of the same nodule, so there was no physical change. There was a striking difference between the thin and thick slice results. For thin slice, the absolute range of reported change measurements was less than 10%. For the thick slice data, the range was about 40%.

A follow-on study to the Biochange '08 pilot is the planned full scale Biochange Challenge. It also uses the RIDER lung CT studies but mainly has thin slice studies, including the MSKCC Coffee Break data discussed earlier. In addition to the participation of algorithm/software developers, the planned study seeks the participation of radiologists to provide markup for comparison with the computed change measures.

A second study group members have conducted is the "VOLCANO'09 Challenge" [32]. This challenge invited participants to evaluate the change in size of pulmonary nodules. The challenge involved measuring the change in nodule size for 50 scan pairs (see example in Fig. 7). Four additional scan pairs were made available for training. The data was selected from cases prepared for the Public Lung Database to Address Drug Response. This database was sponsored by the Prevent Cancer Foundation [33] and provides information on a number of aspects of lesion measuring by means of sample image; this resource is complimentary to the RIDER database. A key component of this database is repeat scans made at the same time. This zero change data set is similar to the No-Change data set except that scans were obtained from the start of CT guided biopsy procedure before the needle affects the image quality. (An example of computer assisted segmentation of the lesions in Fig. 7 is shown in Fig. 8.)

Teams reported the fractional change in nodule size for each of the 50 scan pairs. Thirteen different teams submitted their measurement change results from a total of 17 different methods. In 11 of these cases, the actual volumes recorded for each nodule were also reported. The participants were only informed that there were 50 nodule pairs; however, the data may be divided into four subgroups:

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

A. (14) zero-change in which the scans were taken minutes apart and therefore there is no real change in the nodule size.

B. (13) zero-change cases as in A above except that one scan has a slice thickness of 1.25 mm and the second scan has a larger slice thickness (2.5 or 5.0 mm)

C. (19) nodules with a significant time interval between scans and therefore some real change and (3) nodules with a large amount of size change (greater than 1.5 times in volume). Of these nodules 19 were considered to be stable or benign by biopsy and 3 were diagnosed as malignant.

D. (1) synthetic phantom nodule with a known size recorded with a different slice thickness +

If we only used zero-change data then any system that had a constant output set to zero would be considered to have an ideal response. For this reason we included cases for which a real change was indicated by observation; however, for these cases there is no way to know precisely how much that change is. Most evaluation methods for CAD systems, including challenges, involve a ground truth established be experts. However, for the task of nodule size estimation it is well known that there is a large amount of variation or disagreement in expert size estimations [34]. Further, it has not been established that expert's manual estimations are superior to automated measurements. In this challenge, while the change in size of nodules was reviewed by experts, the issue of ground truth was explored through the submitted responses to the challenge.

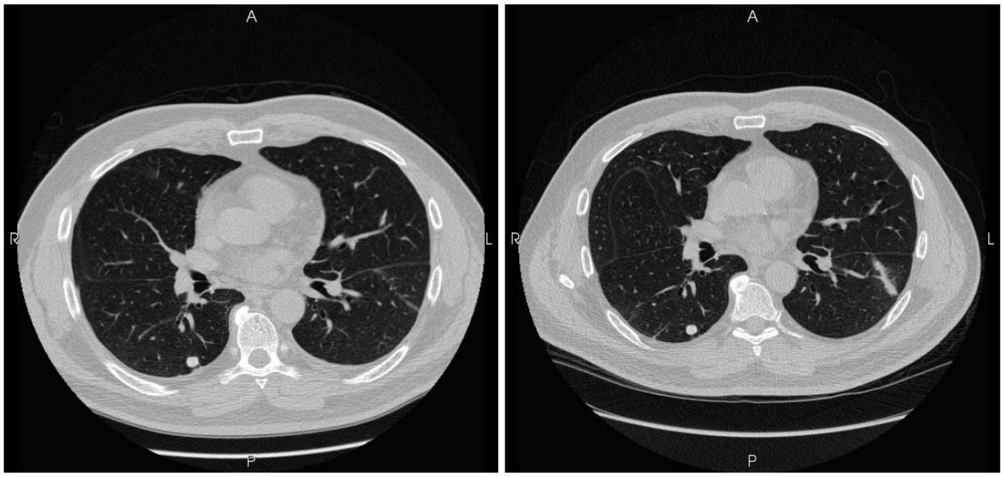


Fig. 7. Two scans of a lesion in the VOLCANO Data set (View 5).

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010

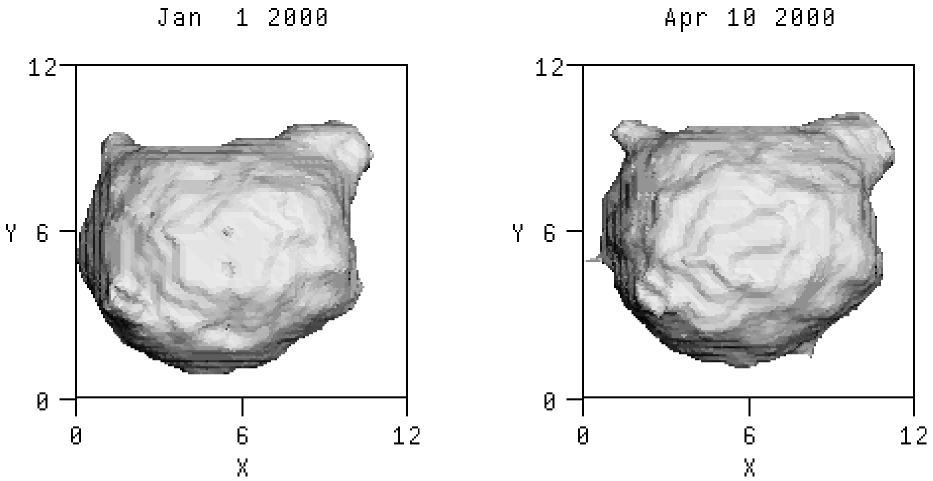


Fig. 8. An example of computer assisted segmentation for the lesions shown in Fig. 7

The initial findings of this study showed there was no statistical difference between the automated methods on scans of the same slice thickness (subgroup A of 14 cases, $p = 0.92$ according to the Friedman test), but there was a statistical difference in the methods when the scan slice thickness is changed (subgroup B of 13 cases, $p < 0.01$ according to the Friedman test). The behavior of the methods for nodules with a small real change in size was similar to that for the zero-change data. The last point has implications for the validity of using zero-size change data sets for evaluating nodule measurement performance. There was an interesting concordance between the different automated methods for a measured change in size for some cases in the zero-change data set. A follow on to this study is VOLCAMAN'10 [35], which enlists a number of physicians using simple manual image marking tools to measure the change in size of the a subset of the cases used in VOLCANO'09. In this way the variation of experts for the same task will be established and comparisons with computer methods can be made.

4. Discussion

These examples are only a small portion of what could be done to advance the field. Whether considered from the vantage point of providing an objective basis on which to evaluate the relative performance of different candidate methods, or to allow individual groups access to larger data sets than they would otherwise be able to afford individually, or as a primary driver in the effort to harness the strength of current and new technology towards clinically relevant problems, there is a recurrent theme of the importance of public data resources. Moreover, the ability to evaluate the same data in different ways is arguably not only helpful, but in fact necessary, to establish an objective basis for performance assessment.

In recognition of the need to improve the availability of public image databases for quantitative imaging research, the RSNA has started an Ad Hoc Committee on Open Image Archives. The main objective of this committee is to make recommendations that have the potential to significantly improve the number, size and quality of open image archives. This will be accomplished by reviewing the history of image archives, identifying the main challenges, incentives, and hurdles to creating such archives, and ultimately create a list of recommendations that will improve open image archives with respect to specific image quantification use cases. It is envisioned that the long-term results of this committee will encourage a new generation of data collections available in open image archives for quantitative imaging.

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010
(C) 2010 OSA

This paper identifies several early programs to collect and utilize data either directly in the public domain or easily accessible to teams that demonstrate their need for it to consortia or other groups that recognize a role in collecting and curating such data. Likewise, it is published using the nascent method referred to by this journal as "interactive science publishing," which further encourages a means by which not only the results but also the data used in deriving those results is available for public peer review. We support the editors position that such capabilities will not only move the state of the art in scientific publication forward, but the science itself will benefit as more access is granted to independent reviewers. Such capability is concordant with the goals of our group and we are pleased to be able to exercise it for our present purposes.

Other working material of the team is maintained on a Wiki page that enables the group activity [36].

Acknowledgments

Many members of our Technical Committee have made substantial contributions to the work described in this report. We would like to specifically acknowledge the following people in addition to the authors (alphabetically): Robert Ford (RadPharm), David Gustafson (Intio), Philip F. Judy (NLST), Matthias Thorn (Siemens Healthcare), James Mulshine (Rush), Daniel Nicolson (Definiens), Kevin O'Donnell (Toshiba), and Daniel C. Sullivan (Duke University and Scientific Advisor to RSNA). Linda Bresolin, Joseph Koudelik, and Fiona Miller of the RSNA organized each meeting, kept our records, and built the contents of our wiki. We would further like to acknowledge the support of the NCI Cancer Imaging Program, its RIDER Project (http://imaging.cancer.gov/reportsandpublications/ReportsandPresentations/LungImaging/), and the CDRH/NIBIB Laboratory for the Assessment of Medical Imaging Systems (LAMIS) (http://www.nibib.nih.gov/Research/Intramural/LAMIS)

#125764 - $15.00 USD Received 22 Mar 2010; revised 11 Jun 2010; accepted 11 Jun 2010; published 2 Jul 2010