Change in lung tumor volume as a biomarker of treatment response: a critical review of the evidence

P. D. Mozley1*, L. H. Schwartz2, C. Bendtsen3, B. Zhao2, N. Petrick4 & A. J. Buckler5

1Department of Imaging, Merck Research Laboratories, West Point, PA; 2Department of Radiology, Columbia University, New York, NY, USA; 3Department of Applied Computer Science and Mathematics, Merck Research Laboratories, Rome, Italy; 4Center for Devices and Radiological Health, USA Food and Drug Administration, Silver Spring, MD and 5Buckler Biomedical, LLC, Boston, MA, USA

Specific Aim: To review the evidence indicating that volumetric image analysis of computed tomography scans meets specifications for qualification as a biomarker in clinical trials or the management of individual patients with lung cancer.

Methods: Claims of value were broken down into questions about technical feasibility, accuracy, the precision of measurement, sensitivity, the correlations with health outcomes, and the risks of producing misleading information. For each claim, the pertinent literature was reviewed.

Results: Technical feasibility has now been shown, but only in limited contexts. Accuracy has been demonstrated, but only for tumors with favorable anatomical features. Measurement error still makes the assessment of change in small nodules precarious in diagnostic settings unless rigorous image acquisition and analysis procedures are followed. Precision is sufficient in some larger masses to make volumetrics a sensitive biomarker. In a few trials, correlations with clinical outcomes have been higher for volumetric-based measures than for unidimensional or bidimensional diameters. Value in clinical practice settings and clinical trials has been suggested, but not proven.

Conclusion: The weight of the evidence indicates there are circumstances in which volumetric image analysis adds value to clinical trial science and the practice of medicine.

Key words: cancer, CT, image analysis, imaging, lung cancer, tomography, computed

INTRODUCTION

RECIST Version 1.1 is a latent standard for serially assessing the longitudinal course of illness in patients with lung cancer [1]. Concerns have been raised about RECIST-based response assessments, in part because tumors do not always expand or contract uniformly, changes in line lengths represent only a small fraction of the available information in the images [2], and the stable disease category is so broad that it is not always adequately sensitive to changes in tumor mass.

In clinical trial settings, this lack of sensitivity can translate into a loss of statistical power per subject enrolled. As a consequence, when all other things are equal, more subjects need to be enrolled in each arm of a trial and each subject who is enrolled needs to remain on study longer. Both effects decrease the number of new compounds that can be tested in clinical trials, increase the costs of drug development, and slow the delivery of new treatments to patients with unmet medical needs.

Analogously, in clinical practice, more sensitive measures of response may be helpful [3]. Individual patients want to know as soon as possible if their new treatments are conveying benefits. If they are not, then they want to launch alternatives as soon as possible. No one wants to take risks for no benefit, or waste time, effort, and money on futile treatments. All stakeholders, including third-party payers, view the problem similarly.

Measuring changes in the volume of an entire tumor could solve some of these problems. However, questions have been raised about whether volumetric image analysis will add value or only increase the costs of patient care and the complexity of running clinical trials. After all, visual inspections of serial computed tomography (CT) scans can be sufficient to demonstrate new metastases that trigger an assessment of treatment failure. Even without findings of new metastases, changes in tumor masses can be so conspicuous that quantification seems pointless. The question, then, is whether there are specific contexts in which volumetric image analysis can be shown to change treatment decisions in ways that convey medically meaningful benefits to individual patients or genuinely enhances the quality of clinical trials. This review sought evidence that indicates volumetric image analysis is or is not an accurate, precise, sensitive, and medically valuable biomarker of response in the assessment of lung tumors.
methods

The Quantitative Imaging Biomarker Alliance (QIBA) [4] constructed a systematic ‘process map’ [5] for eventually qualifying or discarding volumetric CT as a biomarker of response to treatments for a variety of medical conditions, including lung cancer. As part of its due diligence, a QIBA taskforce reviewed the literature to find evidence supporting or refuting the following hypotheses:

- It is technically feasible to measure lung tumor volumes with CT.
- Measurements of these tumor volumes are accurate.
- Most of the technical factors that influence the precision of volumetric measurements are known and can be controlled for.
- The precision, and therefore the sensitivity, of volumetric image analysis for detecting responses is higher than the sensitivity of RECIST.
- The risks of misleading results in clinical trials and the care of individual patients are known and within acceptable limits.
- Thresholds for classifying changes in volume as biologically meaningful can be established with confidence.
- Changes that are greater than these thresholds are medically meaningful and have the potential to be qualified as surrogates for changes in health outcomes.

The literature was reviewed by using PubMed and several Internet search engines. Key words included the following: lung cancer, volume, RECIST, image analysis, outcomes, clinical trials, and some of their synonyms.

results

technical feasibility

The potential value of quantifying tumor volume before and after treatment to assess response was recognized before the introduction of CT [6]. For >30 years now, investigators have argued that the serial quantification of lung mass volume with CT is feasible [7] and can have a positive impact on patient care [8, 9].

Today, many software tools are available for quantifying volume. While expert supervision is still required to prevent these algorithms from ‘getting lost’, ‘running away’, and returning erroneous results, doubts about technical feasibility per se have almost vanished and given way to questions about accuracy, precision, and value in specific settings.

accuracy: phantom studies. In 2000, Yankelevitz et al. [10] showed that the volume of small nodules could be accurately measured to within ±3% provided that their contrast with the surrounding media is high and the image reconstruction interval (RI) is small.

In 2003, Winer-Muram et al. [11] characterized how accuracy was inversely related to the RI. They showed that the smaller the tumor volume to RI, the more the measured volume was overestimated.

In 2008, Ravenel et al. [12] reported that precision and accuracy were highly influenced by slice thickness, but relatively resistant to the reconstruction kernel. They also found that the volumes of small nodules were consistently overestimated. All other things being equal, they confirmed that the larger the object, the higher the precision and accuracy.

In 2009, a group of scientists at the USA Food and Drug Administration began reporting on a series of systematic studies of anthropomorphic phantoms [13, 14]. They confirmed that performance is dependent on slice thickness and object size. Their findings and review of the literature [15] showed that adequately precise and accurate volume estimates are possible, at least when conducted by dedicated scientists working at a single center. However, fundamental questions related to the medical meaning of change still remain to be answered.

accuracy: clinical studies. General proof-of-concept about the ability of CT to accurately quantify volumes in vivo seems best established for solid organs that have subsequently been excised or explanted [16, 17]. However, the confirmation of accuracy in the field of lung cancer remains problematic. Ex vivo volume measurements would require careful dissection of all surrounding tissues. Even if all the normal tissues could be dissected away from a neoplastic mass, maintaining the microcirculatory system, interstitial turgor pressure, intracellular fluid levels, and other physiological state characteristics that influence the true volume of masses in vivo seems as though it would be laborious at best. As a consequence, there are only limited clinical data about the corroboration of accuracy in any type of cancer. It seems likely that accuracy will need to be inferred from images of phantoms, while most clinical imaging research in oncology will need to focus on precision and correlations with health status.

technical factors influencing the precision of clinical volume measurements

There appears to be a consensus that image quality influences precision. Specifically, clinical research studies have shown that precision is (i) inversely proportional to the RI or slice thickness [18, 19]; (ii) directly proportional to the size of the mass [20]; (iii) inversely proportional to the complexity of its shape [21, 22]; (iv) directly proportional to its contrast with surrounding tissue [20, 22]; and (v) dependent on several other miscellaneous factors related to the training of the image analysts [23], software selection [25, 24], and the like.

The clinical literature confirms the phantom studies which have shown that, all else being equal, the precision of measurement increases as image quality increases, and the larger the tumor volume, the lower the variance. This is because most of the measurement error comes from detecting the edges of a tumor on a stack of two-dimensional images. The smaller the picture elements in each slice and the smaller the RI, the greater the certainty about which voxel corresponds to edge of the tumor. All the tumor edges in a stack of images correspond to its surface in three-dimensional space. The smaller the mass, the higher its surface-to-volume ratio, and thus the higher the percent error of measurement. Conversely, the larger the mass, the lower its surface-to-volume ratio and the less susceptible its volume is to measurement error.

sensitivity as a function of precision. The sensitivity of serial measurements for detecting change is highly related to the precision of each measurement. A number of investigators have found that in the context of small lung masses, line length measurements cannot be made with satisfactory precision and can lead to misclassification of response [25–27].
In 2003, Revel et al. [28] studied the intra- and interrater reliability of line lengths in 54 pulmonary nodules that ranged in size from 3 to 18 mm. Agreement was relatively low, leading this team of investigators to conclude that ‘two-dimensional CT measurements are not reliable in the evaluation of small noncalcified pulmonary nodules’. In a 2004 follow-up study, Revel et al. [29] had three raters quantify the volume of these nodules three times each. They found that volume could be quantified in 52 of 54 (96%). Of these 52, there was almost no variation among readers in 35 (67%), and the coefficient of variation for the remaining 17 (33%) averaged 2.26% (range: 2.4%–3.1%). The findings led these investigators to conclude that volume measurements were more reliable than longest diameters in this context. This could be important, because if there is at least one known setting where volumetric image analysis does well, then it might be that volumetrics also outperforms RECIST in other contexts with similar features related to size, shape, contrast, etc.

**Risks of using changes in volume as biomarkers.** Some concerns about using volumes as biomarkers are not related to the technical veracity of measurements, but rather to the ability of the changes to reflect changes in the state of the disease that trigger changes in patient management. For example, in a 2007 review, Shankar et al. [30] noted that RECIST line lengths representing the longest diameter of a mass can (i) underestimate the benefits of targeted therapies that prolong survival despite no visual evidence of tumor shrinkage; (ii) signal misleading indications of disease progression when tumors swell due to bleeding, edema, etc.; and (iii) fail to reflect the appearance of new neoplastic tissues within complex masses.

These problems could also confound measurements of whole tumor volumes. In fact, it is theoretically possible that the added sensitivity of volumetric image analysis could amplify some of these ‘errors’.

In support of the caveats by Shankar et al., a number of investigators have concluded that neither line lengths nor volumes are adequately useful biomarkers of clinical outcome. Some of these reports have been based on studies of heterogeneous tumors in other types of cancer, but could apply to lung tumors as well, particularly if tissue segmentation algorithms are not used to limit the measurements to neoplastic tissues within complex masses.

**Establishing thresholds for classifying response.** Several groups of investigators have reported that currently available image analysis software produces high levels of inter- and intrarater reliability on ‘static’ image sets. However, interscan variability seems much higher when measured in ‘coffee break’ designs. In coffee break studies, patients are rescanned after very short time intervals that require subjects to get off the imaging table after the first scan, and then climb right back on to the table for a second scan. The assumption is that fundamental tumor biology and scanner performance do not change between the first and second scans, even though some factors might, such as patient positioning and inspirational effort.

In 2004, Boll et al. [31] suggested that some interscan differences could be real, and based on true biological changes that occur over extremely short time intervals. They observed that hemodynamic factors can produce true lung nodule compression and expansion. They quantified the volumes of 73 small nodules in 30 patients with a 16-detector CT scanner. They used cardiac gating to show that volume measurements of small nodules vary by as much as 34% during the cardiac cycle. The nodules they studied ranged in size from 0.2 to 399 mm³, corresponding to longest diameters ranging from <1 mm to a little more than 9 mm. However, no one has ever reported biological changes of this magnitude in larger masses.

In 2004, Wormanns et al. [32] published a coffee break study in which they acquired two CT scans of the chest in 10 patients with 50 measurable lung nodules that ranged from 2 to 20 mm in longest diameter. They found that both inter- and intrarater variability in the measured volumes were <1% on any given image. However, interscan variability averaged about ±20%. This could convey some advantages over reliance on longest diameters in this setting.

As referenced above, Gietema et al. [21] reported a coffee break study of 218 lung tumors in 20 patients with metastatic lung cancer. The analyses were limited to masses with a longest diameter of <10 mm. The investigators found that the 95% confidence interval (CI) for differences in measured volumes was 21.2%, 23.8% (mean difference, 1.3%). Although their warning that ‘variation of semiautomated volume measurements of pulmonary nodules can be substantial’ is important in this context, their findings show that volumetric image analysis is feasible and potentially more sensitive than reliance on longest diameters.

In 2009, Zhao et al. [33] reported smaller variances in lung tumor volumes during a coffee break study of 32 patients with advanced lung cancer than most previous studies of small lung nodules. The 95% CI in this study ranged from −12.1% to 13.4%. One factor that might have contributed to the higher precision of measurement in this study was the tumor size, which averaged 3.8 cm in longest diameter. This seems substantially more sensitive than RECIST, and in some settings, might be worth the extra effort required to conduct the analyses.

**Value.** Jaffe [34] pointed out that the value of elegant image analysis has not been proven yet in clinical trials. In the context of this review, value is defined as the ability of imaging to have a meaningful impact on decisions about patient care by predicting the clinical course of illness, or the response to treatment, sooner than alternative methods of assessment.

Suggestions of value are mounting. In 2006, Zhao et al. [35] reported a study of 15 patients with lung cancer at a single center. They used multidetector CT scans with a RI of 1.25 mm to semiautomatically quantify unidimensional longest diameters, bidimensional cross products, and volumes before and after chemotherapy. They found that 11 of 15 (73%) of the patients had changes in volume of ≥20%, while only 1 (7%) and 4 (27%) of the subjects in this sample had changes in uni- or bidimensional line lengths of >20%. There were seven (47%) patients with changes in volume of 30% or more, while there were no patients with unidimensional line length changes of ≥30%, and only two (13%) with changes in bidimensional cross products of ≥30%. The investigators concluded that volumetric image analysis was substantially
more sensitive to drug responses than uni- or bidimensional line lengths in this clinical trial context. However, their initial dataset did not address clinical value in terms of health outcomes.

In a follow-up analysis [36], the same group used volumetric analysis to predict the biological activity of endothelial growth factor receptor (EGFR) modulation in lung cancers with EGFR mutation status as a reference. In a population of 48 patients, changes in tumor volume at 3 weeks after the start of treatment were found to be more sensitive and equally specific to changes in longest diameter for predicting EGFR mutation status. The positive predictive value of early volumetric response for EGFR mutation status in their patient population was 86%. The results were consistent with findings that showed that volumetric image analysis can predict clinical response much sooner than RECIST in other cancers [37]. The investigators concluded that early volume change has promise as an investigational method for detecting the biological activity of systemic therapies in lung cancer.

In 2008, Altorki et al. [38] reported that volumetric image analysis is substantially more sensitive than changes in unidimensional diameters. In a sample of 35 patients with early-stage lung cancer, they found that 30 of 35 (85.7%) subjects treated with pazopanib had a measurable decrease in tumor volume, while only three of these 35 subjects met RECIST criteria for a partial response. However, this group has not yet reported how well decreases in tumor volume corresponded to clinical outcome. Currently, the literature on changes in volume as predictors in long-term health outcomes remains limited.

In 2009, van Klaveren et al. [39] used absolute volumes and doubling times to make diagnostic decisions in 7757 subjects at high risk for lung cancer who were enrolled in the experimental arm of a randomized clinical trial to evaluate the mortality reduction benefit of screening with CT. Patients with lung nodules were followed up for up to 4 years after enrollment. Harmonized image acquisition and analysis protocol were used to produce 1-mm-thick slices at a RI of 0.7 mm. The overall sensitivity of case finding for nodules that met the protocol definition of suspicious was 94.6%, and the negative predictive value was 99.9%. They concluded that serial measurements of volume could spare a substantial fraction of patients with suspicious nodules from invasive diagnostic procedures and their associated morbidity.

**discussion**

There are now many reports describing the feasibility of quantifying lung tumor volumes with CT. Volumetric measurements of solid tumors can be accurate in the proper setting. The precision of measurement is continuously improving, and usually higher than that for corresponding measurements of longest diameter. The sensitivity of volumetrics for distinguishing between measurement error and medically meaningful changes in tumor biology is dependent on context. The literature shows that the context includes areas where there are still intense needs for quantitative measures of response within subjects.

Anthropomorphic phantoms studies and the clinical research literature show that, all else being equal, the uncertainty of measurement decreases as image quality increases, and the larger the tumor volume, the lower the variance. This seems important. In diagnostic settings, distinguishing benign lung nodules from early stages of cancer based on their rates of growth over relatively short intervals is feasible and can spare patients from invasive procedures [39]. However, longitudinal measurements of small nodules require rather rigorous control over the image acquisition and analysis procedures [40]. In contrast, when all other things are equal, larger masses in patients with established diagnoses of lung cancer seem more resistant to measurement error [35]. These relationships between precision and size hold for objects with similar shape. As the geometric shape of a tumor mass becomes increasingly complex, precision tends to decrease. Because spiculation [41] and irregular patterns of growth tend to increase over time, more research is required before the precision of measurement can be succinctly summarized in patients with advanced disease.

It seems hard to overemphasize that, whatever its problems, the literature consistently indicates that volumetric image analysis can be more informative than measurements of line lengths placed on a single tomographic slice. If nothing else, volume measurements obviate the problems that stem from the fact that lung cancers are rarely well modeled as uniformly contracting or expanding spheres.

While it has already been shown that the whole thoracic tumor burden will not be quantifiable in every case, evidence is mounting that volumetric measurements can, and should, influence treatment decisions in patients with lung tumors.

It seems likely that volumetrics will also succeed in other types of extra-thoracic cancer when the tumors are the right size, shape, and density when compared with the surrounding tissue. Although there is not yet enough evidence to claim that volumetric image analysis is qualified as a biomarker of response in patients with solid tumors, quantifying changes in tumor volume could constitute a major paradigm shift in clinical practice, as well as the conduct of some clinical trials.

**acknowledgements**

QIBA’s Volumetric CT Committee contributed heavily to the formation of the hypotheses and critical concepts.

**disclosures**

There have been no involvements that might raise the question of bias in the work reported or in the conclusions, implications, or opinions stated.

**references**


