Radiology Advances, 2025, **2(1)**, umaf001
https://doi.org/10.1093/radadv/umaf001
Advance access publication: January 6, 2025
Original Research

Comparing quantitative imaging biomarker alliance volumetric CT classifications with RECIST response categories

Binsheng Zhao, DSc*[,1], Nancy Obuchowski, PhD[2], Hao Yang, MS[1], Yen Chou, MD[3], Hong Ma, MD[4], Pingzhen Guo, MD[1], Ying Tang, PhD[5], Lawrence Schwartz, MD[1], Daniel Sullivan, MD[6]

[1]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, United States
[2]Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH 44195, United States
[3]Department of Radiology, Fu Jen Catholic University Hospital, New Taipei City 24352, Taiwan
[4]Department of Radiology, Columbia University Irving Medical Center, New York, NY 10032, United States
[5]Department of Clinical Research and Regulatory Affairs, CCS Associates, McLean, VA 22102, United States
[6]Department of Radiology, Duke University Medical Center, Durham, NC 27710, United States

*Corresponding author: Binsheng Zhao, DSc, Memorial Sloan Kettering Cancer Center, 321 East 61st Street, Rm 232, New York, NY 10065, United States (zhaob1@mskcc.org).

Abstract

Purpose: To assess agreement between CT volumetry change classifications derived from Quantitative Imaging Biomarker Alliance Profile cut-points (ie, QIBA CTvol classifications) and the Response Evaluation Criteria in Solid Tumors (RECIST) categories.

Materials and Methods: Target lesions in lung, liver, and lymph nodes were randomly chosen from patients in 10 historical clinical trials for various cancers, ensuring a balanced representation of lesion types, diameter ranges described in the QIBA Profile, and variations in change magnitudes. Three radiologists independently segmented these lesions at baseline and follow-up scans using 2 software tools. Two types of predefined disagreements were assessed: Type I: substantive disagreement, where the disagreement between QIBA CTvol classifications and RECIST categories could not be attributed to the improved sensitivity of volumetry in detecting changes; and Type II: disagreement potentially arising from the improved sensitivity of volumetry in detecting changes. The proportion of lesions with disagreements between QIBA CTvol and RECIST, as well as the type of disagreements, was reported along with 95% CIs, both overall and within subgroups representing various factors.

Results: A total of 2390 measurements from 478 lesions (158 lungs, 170 livers, 150 lymph nodes) in 281 patients were included. QIBA CTvol agreed with RECIST in 66.6% of interpretations. Of the 33.4% of interpretations with discrepancies, substantive disagreement (Type I) occurred in only 1.5% (95% CI: [0.8%, 2.1%]). Factors such as scanner vendor ($P = .584$), segmentation tool ($P = .331$), and lesion type ($P = .492$) were not significant predictors of disagreement. Significantly more disagreements were observed for larger lesions (≥50 mm, as defined in the QIBA Profile).

Conclusion: We conclude that QIBA CTvol classifications agree with RECIST categories.

Keywords: tumor volumetry, volumetric response assessment, response assessment criteria, computed tomography

Abbreviations
QIBA = quantitative imaging biomarker alliance; CTvol = volumetric CT; RECIST = response evaluation criteria in solid tumor; Vol-PACT = advanced metrics and modeling with volumetric CT for precision analysis of clinical trial results; PR = partial response; SD = stable disease; PD = progressive disease; CR = complete response; wCV = within-subject coefficient of variation.

Summary
Agreement between Quantitative Imaging Biomarker Alliance (QIBA) volumetric CT (CTvol) and RECIST classifications was established, which facilitates the use of precise volumetry for tracking tumor changes for treatment assessment.

Key Results
- QIBA CTvol classifications show strong agreement with the RECIST category system.
- QIBA CTvol classifications demonstrate a potential advantage over RECIST-recommended volume cut-offs in detecting tumor response and progression when utilizing volume measurement.

Introduction

In the era of precision medicine, targeted, immune, and combination cancer therapies are rapidly advancing, with the goal of improving treatment outcomes. While clinical outcomes, such as overall survival (OS), continue to be the gold standard for assessing the value of new drugs, reaching these outcomes in clinical trials may require years and large numbers of patients. For over 2 decades, Response Evaluation Criteria in Solid Tumors (RECIST) have served as a foundational concept in oncology trials, using change in tumor

Received: September 11, 2024; **Revised:** November 27, 2024; **Accepted:** December 30, 2024

unidimensional measurement obtained from serial computed tomography (CT) scans as a surrogate marker. RECIST employs a 30% decrease cut-off to define tumor response and a 20% increase cut-off to identify tumor progression.[1,2]

Evidence indicates that higher-dimensional and potentially more clinically informative data, such as tumor volume and more comprehensive radiomic features, can be accurately and practically quantified by contemporary scanners using advanced image analysis methods.[3] Some studies have shown that the CT volumetric technique demonstrated the potential to serve as an early and more precise biomarker for drug development. For example, a pilot study found that three weeks post-gefitinib therapy, tumor volume change was more closely correlated with the presence of a sensitizing *EGFR* mutation than changes in unidimensional measurements in early-stage non-small cell lung cancer (NSCLC) patients.[4] According to a colorectal cancer (CRC) study, combining tumor volume measurement with tumor growth rate estimated using exponential growth modeling revealed enhanced detection of treatment effects of aflibercept or panitumumab added to standard chemotherapy, surpassing what can be achieved with RECIST unidimensional measurement.[5]

Despite promising findings, the lack of clear guidance on using tumor volumetry as a response assessment biomarker hinders its widespread validation and adoption. Recently, controversial results have emerged regarding the clinical superiority of volumetry over unidimensional techniques, primarily due to the use of varying response cut-offs for volume assessments.[6–11] For example, the RECIST-suggested volume cut-offs for defining response and progression are −65% and +73%, respectively, which correspond to unidimensional changes of −30% and +20% for a spherical tumor that changes symmetrically. Although high-resolution imaging and advanced segmentation software allow for accurate volume measurement and extensive research has been conducted on measurement reproducibility,[12] there remains no consensus on response cut-off values for volumetric assessment.

After years of discussions among experts and analysis of relevant scientific publications on measurement reproducibility, the Quantitative Imaging Biomarkers Alliance (QIBA) has published its Profile "CT Tumor Volume Change for Advanced Disease (CTV-AD)."[13] The QIBA Profile claims that: "A true change in tumor volume has occurred with 95% confidence if the measured volume change is larger than 24%, 29%, or 39% when the longest in-plane diameter at baseline is within 50-100 mm, 35-49 mm, or 10-34 mm, respectively." The claims in the QIBA Profile regarding the within-subject coefficient of variation (wCV) and cut-points for determining the presence or absence of real volume change were based on published test–retest repeatability studies. The primary objective of this study was to evaluate the agreement between the QIBA Profile's CT volumetry biomarker definition of change (ie, QIBA CTvol classifications, QIBA CTvol for short) and the RECIST unidimensional category system (RECIST for short) of partial response (PR), complete response (CR), stable disease (SD), and progressive disease (PD).

Materials and methods

Study overview

This study utilized fully de-identified CT images obtained from a previous study, advanced metrics, and modeling using volumetric CT for precise analysis of clinical trial outcomes, known as Vol-PACT.[14] In the authors' institutions, investigations utilizing external, de-identified patient imaging datasets with no associated link to protected health information are considered non-human subjects research and thus exempt from institutional review board oversight. Figure 1 illustrates the study workflow. Subsequent sections provide further detail.

Patient and lesion data collection

Retrospectively, patient images were obtained from 10 historical Phase III drug trials collected in Vol-PACT. During the Vol-PACT project, target lesions were volumetrically segmented for all CT scan timepoints in each clinical trial, with their anatomical locations recorded. Our analysis focused solely on three types of tumors located in the lung, liver, and lymph nodes, given their prevalence as common sites of metastatic spread. We developed an algorithm (see Supplemental Materials S1 for details) to automate the selection process for patients, images, and lesions, ensuring a balanced representation of lesion types, size ranges, and percentage changes while adhering to QIBA's criteria. Lesion diameters were categorized into small (10-34 mm), medium (35-49 mm), and large (50-100 mm) groups. The magnitude of diameter change was classified as follows: $<-50\%$, -50% to -20%, -20% to $+20\%$, $+20\%$ to $+50\%$, and $>+50\%$. To assist radiologists who were blinded to the Vol-PACT segmentation results, a 1-cm lesion circle marker was automatically placed at the center of each selected lesion on the original image, which had the largest area.

Lesion segmentation

Three radiologists, R1, R2, and R3, with different reading skills (25+, 6, and 11 years of experience in CT interpretation, respectively) and training backgrounds were recruited as independent readers. Two semi-automated lesion segmentation software tools were used: active contour-based segmentation on a customized Weasis platform[15] and GrowCut segmentation in 3D Slicer.[16] There were 2 reading sessions, and there was at least a 2-week washout period between reading sessions of the same cases with different image analysis tools. Computer-generated lesion contours on baseline and follow-up scans were reviewed and modified if deemed suboptimal by radiologists. This process was carried out in a side-by-side fashion to improve segmentation consistency between the two scans. During this procedure, radiologists were advised to use the window/level (W/L) settings predefined for different lesion types. Despite this guidance, radiologists had the flexibility to adjust W/L settings according to their preferences for improved lesion visualization. After completing the segmentation of a lesion, both its longest in-plane diameter (in mm) and volume (in mm^3) were automatically computed from the segmentation mask.

Statistical analysis

The primary objective of this study was to assess agreement between the QIBA CTvol and the RECIST system using the categories of PR, SD, or PD. Table 1 illustrates the a priori definitions of agreement based on the count of lesion measurements, where Δ represents the measured volume change is defined as (follow-up volume − baseline volume) / (baseline volume). The RECIST cut-offs of −30% and +20% were used for the unidimensional measurements, and the

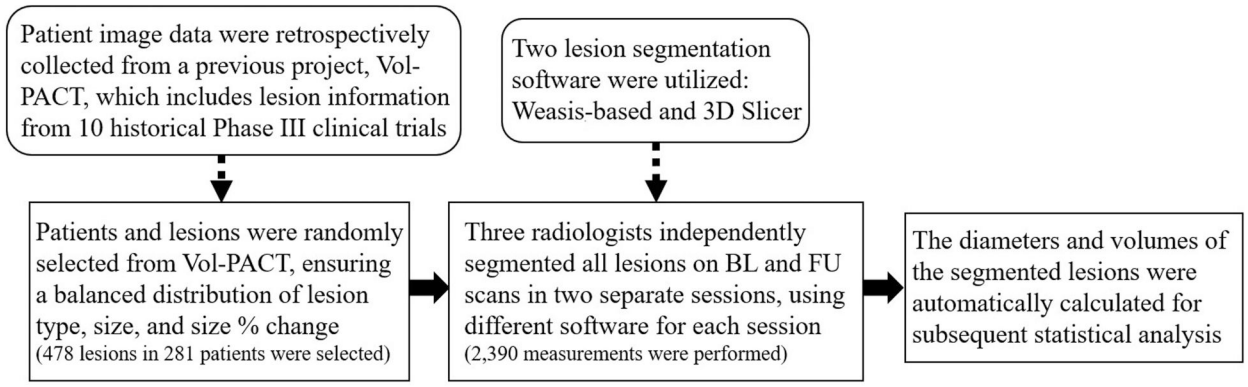

Figure 1. Overview of study workflow.

Table 1. Agreement endpoint definitions.

RECIST categories:	QIBA CTvol classifications[a]		
	$\Delta < x_1$ **"Response"**	$x_1 < \Delta < x_2$ **"Stable"**	$\Delta > x_2$ **"Progression"**
PR or CR (30% decrease in unidimensional measurement)	*Agreement*	*Substantive Disagreement*	*Substantive Disagreement*
SD	*No Agreement*	*Agreement*	*No Agreement*
PD (20% increase in unidimensional measurement)	*Substantive Disagreement*	*Substantive Disagreement*	*Agreement*

[a]x_1 and x_2 are defined from the QIBA profile as −/+39%, −/+29%, and −/+24%, for baseline lesions of 10-34, 35-49, and 50-100 mm diameters, respectively. For example, for small lesions 10-34 mm in diameter, response would be defined as a change in volume from baseline < −39% (denoted as x_1), stable would be defined as a difference between −39% (denoted as x_1) and +39% (denoted as x_2), and progressive defined as a change > +39% (denoted as x_2). For medium baseline lesions (35-49 mm), x_1 becomes −29% and x_2 becomes +29%. For large baseline lesions (50-100 mm) x_1 becomes −24% and x_2 becomes +24%.
Abbreviations: CTvol = CT volumetry, QIBA = Quantitative Imaging Biomarker Alliance.

volumetric response cut-offs of x_1/x_2 were defined from the QIBA Profile CTvol classifications as −/+39%, −/+29%, and −/+24%, for baseline lesions of 10-34, 35-49, and 50-100 mm diameters, respectively. Disagreements between the 2 approaches were further categorized as (1) Type I: substantive disagreement, where the disagreement between QIBA CTvol and RECIST cannot be attributed to the improved sensitivity of volumetry in detecting changes and is undesirable, and (2) Type II: disagreement potentially due to the improved sensitivity of volumetry in detecting changes. The rationale for defining the 2 types of disagreement can be found in Supplementary Materials S2. A third scan (ie, a second post-therapy timepoint scan), when available, was collected and used to further explore these disagreements.

The proportion of interpretations with disagreement between QIBA CTvol and RECIST, as well as the type of disagreement, was reported, along with 95% CIs overall and by sub-groups. The primary null hypothesis was that the proportion of lesions with substantive disagreement is ≥15%; the alternative hypothesis was that the proportion of lesions with substantive disagreement is <15%. The null hypothesis was tested based on the pooled data (across all readers, anatomic locations, and lesion characteristics). The proportion of lesions with substantive disagreement was calculated as the number of lesions categorized as substantive disagreement divided by the total number of lesions. A generalized linear model with generalized estimating equations to account for the clustered data was used to construct a 95% CI for the proportion of substantive disagreement. If the upper 95% confidence bound was <15%, we concluded that these are comparable methods with negligible outliers.

Several factors may affect the agreement between QIBA CTvol and RECIST. Therefore, multiple-variable logistic regression models were built to identify predictors of any disagreement between the 2 classification systems. The independent variables included the scanner vendor, lesion location, segmentation software, and baseline size.

Sample size considerations

The following assumptions were made in determining sample size for the primary study objective: the correlation between measurements on the same lesion is 0.5 (moderate); on average, patients have 2 eligible lesions; 3 readers will make the measurements for all lesions with 2 image segmentation software tools; the proportion of substantial disagreement is between 0.05 and 0.10; and a study with 80% power and 5% type I error rate is required. Based on these assumptions, a study with 234 subjects and 3 readers was proposed, assuming the proportion of substantive disagreement is ≤0.10.

Results

Patient and lesion data

In total, 478 lesions from 281 patients were selected at two time points (median time interval: 68.5 days, range: 15-399 days) from 10 phase III clinical trials (3 CRC, 2 renal cell carcinomas [RCC], 3 NSCLC, and 2 melanoma) collected through the Vol-PACT project. There were 170 liver metastases (36%), 158 lung lesions (33%), and 150 lymph nodes (31%) distributed from the lung apex to the base of the pelvis. Additional details, including lesion distributions, can be found in Supplementary Materials S3.

Radiologists 1 and 2 provided measurements utilizing the two software tools. However, radiologist 3 only reported measurements using a single image analysis software (Weasis) due to the difficulties encountered while working with 3D Slicer (see Discussion). The analysis thus incorporated a total of 2390 measurements, with 956 measurements contributed by radiologist 1, 956 by radiologist 2, and 478 measurements by radiologist 3.

Agreements between QIBA CTvol and RECIST

Table 2 summarizes the pooled overall agreement between QIBA CTvol and RECIST. The QIBA CTvol agreed with RECIST in 66.6% (1592/2390) of interpretations and disagreed with RECIST in 33.4% (798/2390) of interpretations, categorized as (1) substantive disagreement in 1.5% (35/2390) and (2) disagreement potentially due to improved response assessment sensitivity with CT volumetry in 31.9% (763/2390). Since the substantive disagreement, with the 95% CI for the proportion of interpretations of [0.8%, 2.1%], was <15%, the null hypothesis was rejected, and we concluded that QIBA CTvol agreed with RECIST.

To verify SD as determined by RECIST while also presenting a Response or Progression according to QIBA CTvol, we examined the second follow-up scans. At this timepoint, we found that 195 out of 540 lesions (36.1%) initially classified as SD by RECIST and showing a Response by QIBA CTvol were reclassified as a PR by RECIST. Similarly, 101 out of 223 lesions (45.3%) initially categorized as SD by RECIST but displaying Progression by QIBA CTvol were identified as PD by RECIST. Additional details can be found in Supplementary Materials S4.

Agreements between RECIST-suggested CTvol and RECIST

Table 3 outlines the pooled overall agreement between RECIST CTvol and RECIST. RECIST CTvol agreed with RECIST in 86.1% (2058/2390) of interpretations and disagreed with RECIST in 13.9% (332/2390) of interpretations, broken down as (1) substantive disagreement in 6.6% (157/2390) and (2) disagreement potentially due to improved sensitivity with CT volumetry in 7.3% (175/2390). The 95% CI for the proportion of interpretations with substantive disagreement is [5.4%, 8.1%].

Subgroup analyses: effects of variables on agreements between QIBA CTvol and RECIST

In secondary analyses, models were built to identify predictors of disagreement between QIBA CTvol and RECIST. The proportion of substantial disagreement was small for all subgroups analyzed (See Table 4), far less than the hypothesized 15%. Scanner vendor ($P = .584$), lesion segmentation tool ($P = .331$), and lesion anatomical location ($P = .492$) were not significant predictors of disagreement. The only significant predictor of substantial disagreement was baseline size of the lesion, with significantly more disagreements for larger lesions (23.4% disagreement for small lesions, 39.2% for moderately-sized lesions, and 44.2% for large nodules, $P < .001$), though the proportion of substantial disagreement remained low for all lesion sizes.

Lastly, a model was built to assess the effect of the magnitude of change as a predictor of substantial disagreement between the QIBA CTvol and RECIST. The magnitude of change in volume from baseline was defined as large (>50%), moderate (20%-50%), and small (<20%). Among lesions with a large change in volume from baseline, the QIBA CTvol and RECIST substantially disagreed in only 0.1% (2/1454), which was significantly less than for lesions with a moderate or small change in volume (3.3% [19/582] and 4.0% [14/354], respectively) ($P < .004$). Example lesions are shown in Figure 2.

Variability in lesion measurements

Figure 3 plots the inter-reader wCVs by software tool (1) and lesion type (2). The Weasis %wCV estimates were based on 1434 pairs of baseline lesion measurements (478 lesions × 3 combinations of reader pairs); the 3D Slicer %wCV estimates were based on 478 pairs of baseline lesion measurements (478 lesions with just one combination of reader pair). The inter-reader reproducibility did not show differences between the 2 segmentation software tools for both unidimensional and volumetric measurements. However, lymph nodes, particularly when measured using Weasis, demonstrated notably superior reproducibility compared to lung and liver lesions for both unidimensional and volumetric measurements.

Discussion

There is an ongoing need in cancer clinical trials for more accurate and early imaging biomarker, such as tumor volume changes estimated from standard follow-up CT images, to monitor tumor progression. To accelerate the evaluation of volumetric techniques in tumor response assessment, we thoroughly assessed the agreement between QIBA CTvol classifications and RECIST categories using a randomly selected subset of image data from Vol-PACT. Three

Table 2. Agreement between QIBA CTvol and RECIST.

RECIST categories:	QIBA CTvol classifications[a]		
	$\Delta < x_1$ "Response"	$x_1 < \Delta < x_2$ "Stable"	$\Delta > x_2$ "Progression"
PR or CR (30% decrease in unidimensional measurement)	637 (26.7%)	1 (0.0%)	0 (0.0%)
SD	540 (22.6%)	519 (21.7%)	223 (9.3%)
PD (20% increase in unidimensional measurement)	6 (0.3%)	28 (1.2%)	436 (18.2%)

[a] x_1 and x_2 are defined from the QIBA profile as −/+39%, −/+29%, and −/+24%, for baseline lesions of 10-34, 35-49, 50-100 mm diameters, respectively. For example, for small lesions 10-34 mm in diameter, response would be defined as a change in volume from baseline < −39% (denoted as x_1), stable would be defined as a difference between −39% (denoted as x_1) and +39% (denoted as x_2), and progressive defined as a change > +39% (denoted as x_2). For medium baseline lesions (35-49 mm), x_1 becomes −29% and x_2 becomes +29%. For large baseline lesions (50-100 mm) x_1 becomes −24% and x_2 becomes +24%.

Abbreviations: RECIST = Response Evaluation Criteria in Solid Tumors, PR = Partial Response, CR = Complete Response, SD = stable disease, PD = progressive disease.

Table 3. Agreement between RECIST CTvol and RECIST.

RECIST categories:	RECIST-suggested volume cut-offs[a]		
	$\Delta < -65\%$ "Response"	$-65\% < \Delta < 73\%$ "Stable"	$\Delta > 73\%$ "Progression"
PR or CR (30% decrease in unidimensional measurement)	577 (24.1%)	61 (2.6)	0 (0.0%)
SD	104 (4.4%)	1107 (46.3%)	71 (3.0%)
PD (20% increase in unidimensional measurement)	1 (0.0%)	95 (4.0%)	374 (15.5%)

[a]The RECIST-suggested volume cut-offs for defining PR and PD are −65% and +73%, respectively. These correspond to unidimensional changes of −30% and +20% for a spherical tumor that changes symmetrically.
Abbreviations: RECIST = Response Evaluation Criteria in Solid Tumors, PR = Partial Response, CR = Complete Response, SD = stable disease, PD = progressive disease.

Table 4. Substantial disagreements by lesion location, baseline size, software, and scanner.

Variables	% Substantial disagreement [95% CI]
Lesion location	
Liver ($N = 850$)	0.8% [0.1%, 1.5%]
Lung ($N = 790$)	1.8% [0.6%, 3.0%]
Lymph nodes ($N = 750$)	1.9% [0.3%, 3.5%]
Baseline size	
Small ($N = 1214$)	2.1% [1.1%, 3.3%]
Medium ($N = 561$)	1.1% [0.1%, 2.7%]
Large ($N = 615$)	0.5% [0.1%, 1.1%]
Software	
3D slice ($N = 956$)	1.5% [0.6%, 2.3%]
Weasis ($N = 1434$)	1.5% [0.7%, 2.3%]
Scanner	
GE ($N = 665$)	1.4% [0.3%, 2.5%]
Philips ($N = 405$)	1.0% [0%, 2.3%]
Siemens ($N = 705$)	1.4% [0.2%, 2.7%]
Toshiba ($N = 270$)	1.9% [0.3%, 3.5%]

Abbreviation: N = number of lesion measurements.

radiologists independently segmented the three most common types of lesions using two different software platforms.

When comparing QIBA CTvol with RECIST, we found a good agreement (66.6%) between the two systems, with substantive disagreement being extremely low (1.5%). The Stable category in response assessment is crucial clinically, as broadening its range can potentially delay the timely identification of PR or/and PD. If QIBA CTvol classifies a lesion as stable while RECIST classifies it as PR or PD, it suggests that QIBA CTvol may have lower sensitivity in detecting tumor changes compared to RECIST. Only 29 measurements (1.2%) fell into this category. Conversely, if RECIST classifies a lesion as SD while QIBA CTvol classifies it as response or progression, it may indicate that QIBA CTvol has higher sensitivity for detecting changes compared to RECIST. A total of 763 lesions (31.9%) fell into this category. To assess the potential for improved sensitivity in detecting change, we reclassified RECIST's Stable lesions using measurements from the second follow-up scan time point. As shown in Table S2 of Supplementary Materials S4, a notable proportion (195 out of 540; 36.1%) initially classified as Stable by RECIST were subsequently reclassified as PR, along with 101 out of 223 (45.3%) reclassified PD cases, consistent with the findings from QIBA CTvol. Our data thus suggest a potentially higher sensitivity of QIBA CTvol compared to RECIST in detecting PR and PD.

Subsequently, we compared RECIST CTvol with RECIST. Ideally, if lesions are spherical and change symmetrically, there should be perfect agreement between these 2 classification systems. However, our data did not support this assumption, despite our observation of a strong agreement (86.1%), with a low percentage for both types of disagreements (all <7.5%). Previous studies have also reported that RECIST-suggested volumes overestimated actual lesion volume measurements,[17,18] showed discordant response assessment results in 20% of lesions,[19] and tended to more frequently classify the response as PD compared to assessments based on the actual tumor volume.[20]

Compared to RECIST CTvol, QIBA CTvol exhibited much lower agreement with RECIST for the Stable category (519 vs 1107). This is not surprising given the broad range of the Stable category as defined by RECIST CTvol that is extrapolated from unidimensional RECIST. However, QIBA CTvol showed an increase in agreement with RECIST for both PR (637 vs 577) and PD (436 vs 374). Moreover, QIBA CTvol demonstrated a considerable decrease in substantive disagreement (1.5% vs 6.6%) and an increase in type II disagreement (31.9% vs 7.3%). All of the above suggest a potentially greater sensitivity of QIBA CTvol over RECIST CTvol in identifying Response and Progression.

The potential wide range of the RECIST CTvol Stable category may be attributed to its derivation from the Stable category of the RECIST unidimensional system. As is well known, RECIST cut-offs have not yet been proven biologically sensitive or validated by measurement reproducibility in modern times since their establishment.[21,22] Studies reported stronger clinical correlations when response cut-offs lower than those of RECIST were applied to unidimensional measurements.[23,24] For instance, it was reported that a response cut-off value of 20% for early tumor shrinkage (ETS) correlated with longer progression-free survivaland OS in patients with KRAS wild-type metastatic colorectal cancer (mCRC) treated with chemotherapy combined with cetuximab.[23] In another study involving metastatic gastrointestinal stromal tumor patients treated with imatinib mesylate, it was found that responders, identified by either a $\geq 10\%$ reduction in tumor diameter or a $\geq 15\%$ decrease in tumor density, showed a sensitivity of 97% and a specificity of 100% in detecting PET responders. This contrasts with the 52% sensitivity and 100% specificity observed with RECIST. Furthermore, good responders on 2-month CT had significantly longer time-to-progression than non-responders.[25]

This study has several limitations. First, in the absence of clinical outcome data, we relied on the RECIST category system as the reference for the comparisons. As we know, the RECIST system may require some periodic review to ensure its continued applicability with contemporary drugs, modern high resolution imaging devices, and advanced computer-

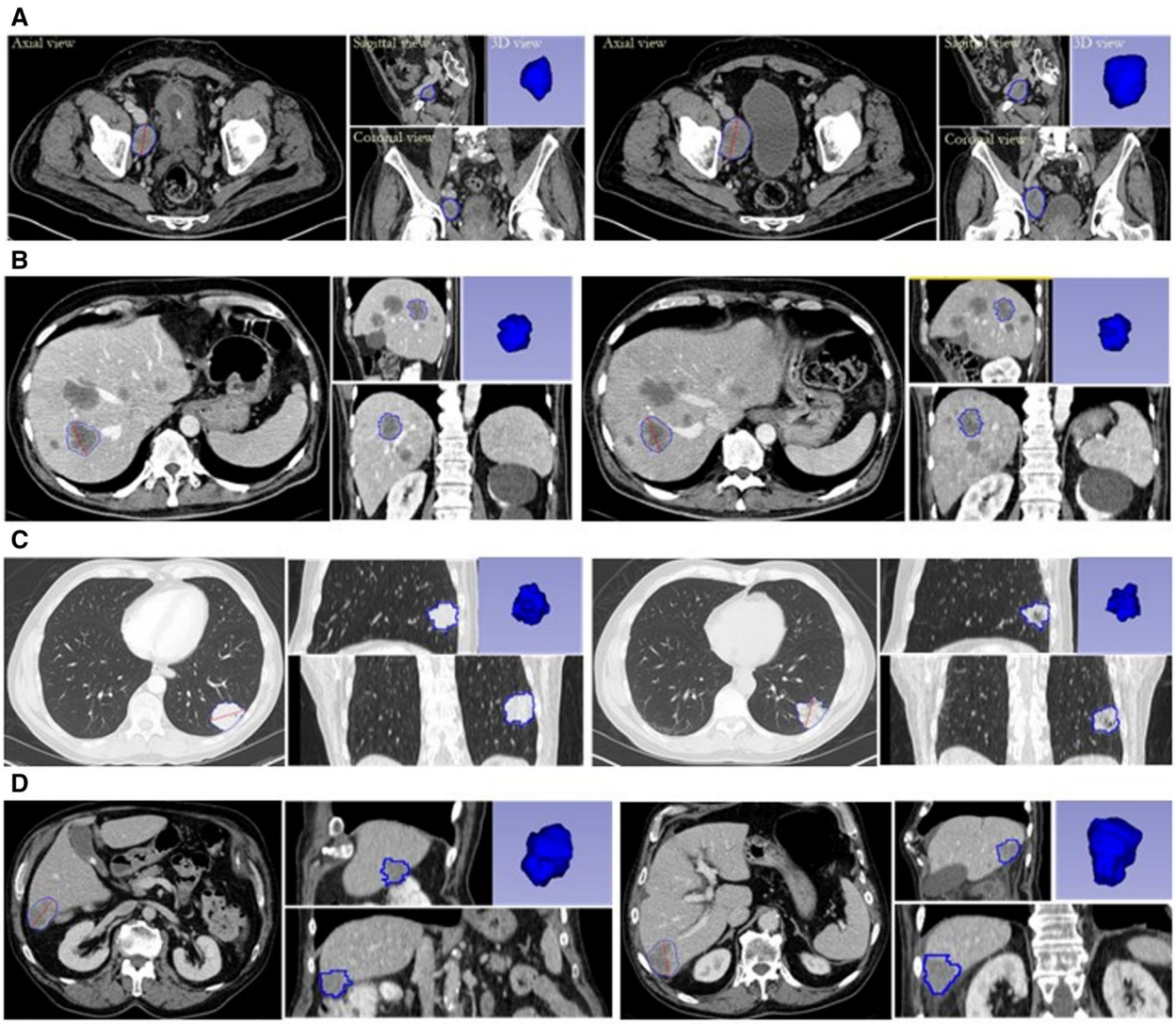

Figure 2. Examples showing agreement and disagreement between QIBA CTvol and RECIST. (A) A pelvic lymph node measured 37.1 mm in diameter and 19 734.2 mm^3 in volume at baseline, and 48.9 mm and 45 715.3 mm^3 at the 3-month follow-up scan. The changes in diameter and volume were +31.6% and 131.7%, respectively. The lesion was classified as progression by both categorization systems. (B) A liver lesion measured 39.4 mm in diameter and 27 012.9 mm^3 in volume at baseline, and 39.6 mm and 23 434.9 mm^3 at the 8-week follow-up scan. The changes in diameter and volume were +0.5% and −13.2%, respectively. The lesion was classified as Stable by both categorization systems. (C) A lung lesion measured 46.6 mm in diameter and 36 311.9 mm^3 in volume at baseline, and 42.6 mm and 24 792.2 mm^3 at the 3-month follow-up scan. The changes in diameter and volume were −8.7% and −31.7%, respectively. The lesion was classified as Stable by RECIST and Response by QIBA CTvol. Notably, a decrease in density over time was observed in this lesion, suggesting it was likely responsive. (D) A liver lesion measured 46.7 mm in diameter and 23 012.7 mm^3 in volume at baseline, and 45.7 mm and 38 886.7 mm^3 at the 8-week follow-up scan. The changes in diameter and volume were −2.1% and 69.0%, respectively. The lesion was classified as Stable by RECIST and Progression by QIBA CTvol.

aided quantification software tools. However, the validation of the RECIST category system falls outside the scope of this study. Second, towards the end of this study, we realized that our dataset included a small portion of RECIST-defined non-measurable lesions (eg, lung lesion embedded in atelectasis), which could have contributed to increased measurement variability, particularly for larger lesions. However, scanner vendor, segmentation tool, and lesion type were not found to be significant predictors of disagreement. Lesions were measured on baseline and follow-up scan images using a side-by-side model, which helped minimize variability in change measurements. Due to page limitations, a detailed discussion of measurement variability was not included in this work. Third, although this study was thoughtfully designed, some unanticipated events occurred during its course. For instance, we underestimated the challenge for radiologists to use the 3D Slicer software platform despite the fact that it is widely used as a research tool. Two radiologists with no prior experience with 3D Slicer encountered difficulties in segmenting, editing, and saving lesions using this software. One of them had to abandon the use of the 3D Slicer for segmentation.

In conclusion, our study found a strong agreement between QIBA CTvol classifications and RECIST category system, which is the current standard for tumor response assessment in clinical trials, and the potential advantage of QIBA CTvol classifications over RECIST-suggested volume cut-offs in detecting tumor response and progression when utilizing volume measurements. Although we established agreement between QIBA CTvol classifications and RECIST response categories, clinical validation is warranted before the QIBA

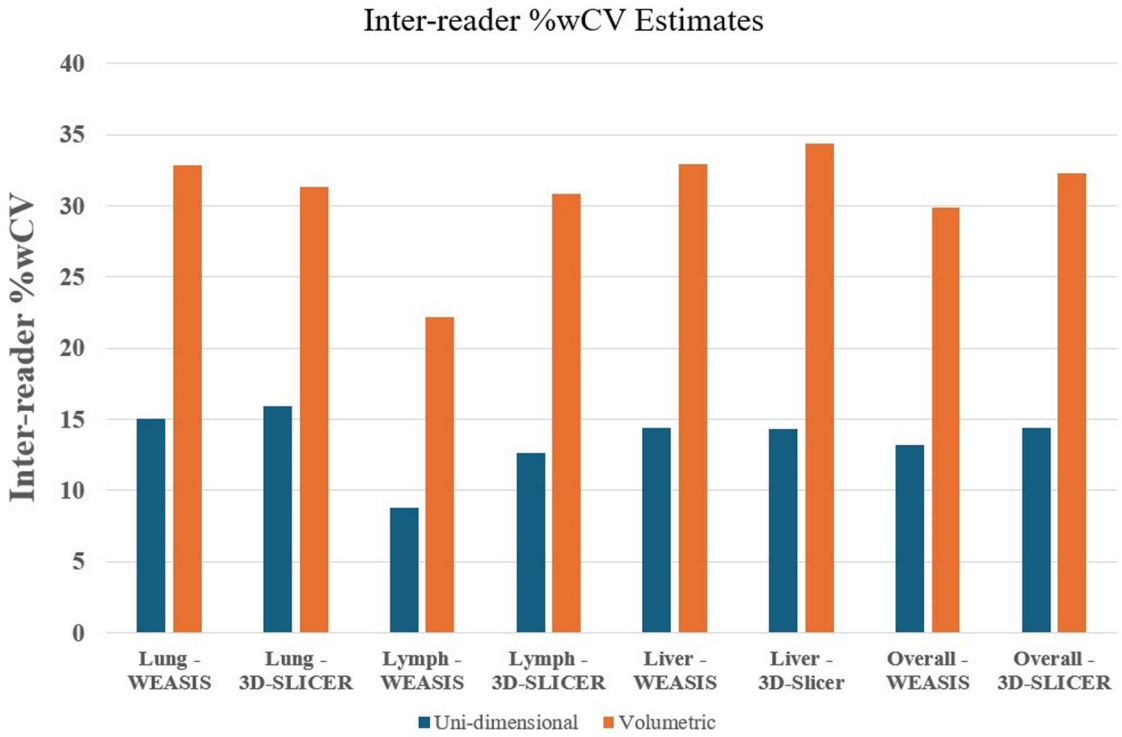

Figure 3. Inter-reader percent within-subject coefficient of variation (%wCV) estimates by software tool and lesion type. The within-subject coefficient of variation is a metric of precision defined as the standard deviation of the measurements (referring here to the SD of the readers' measurements of the same lesion) divided by the magnitude of the measurement (eg, mean of the readers' measurements of the lesion).[13]

CTvol classifications can facilitate the widespread use of precise CT volumetry. Reproducible CT volumetry is crucial for tracking tumor changes, improving treatment assessment, and guiding clinical decision-making.

List of contributors

Binsheng Zhao and Nancy Obuchowski contributed equally to this work.

Author contributions

Binsheng Zhao (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision), Nancy Obuchowski (Conceptualization, Formal analysis, Investigation, Methodology, Software, Supervision, Validation), Hao Yang (Data curation, Methodology, Software), Yen Chou (Data curation), Hong Ma (Data curation), Pingzhen Guo (Data curation), Ying Tang (Conceptualization, Funding acquisition, Investigation, Project administration), Lawrence Schwartz (Conceptualization, Data curation, Investigation, Methodology, Supervision), and Daniel Sullivan (Conceptualization, Investigation, Methodology)

Supplementary material

Supplementary material is available at *Radiology Advances* online.

Funding

This Research was supported in part through the National Institutes of Health (U01 FD007470-01) and the NIH/NCI Cancer Center Support Grant (P30 CA008748). The content is solely the responsibility of the authors and does not necessarily represent the views of the funding sources.

Conflicts of interest

Please see ICMJE form(s) for author conflicts of interest. These have been provided as supplementary materials. All authors have no conflicts of interest.

Data availability

The data utilized in this study were provided as part of an academic industrial partnership for which general data sharing was not contractually allowed.

References

1. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst*. 2000;92(3):205-216.
2. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228-247.
3. Funingana IG, Piyatissa P, Reinius M, McCague C, Basu B, Sala E. Radiomic and volumetric measurements as clinical trial endpoints-a comprehensive review. *Cancers (Basel)*. 2022;14(20):5076.
4. Zhao B, Oxnard GR, Moskowitz CS, et al. A pilot study of volume measurement as a method of tumor response evaluation to aid biomarker development. *Clin Cancer Res*. 2010;16(18):4647-4653.
5. Maitland M, Wilkerson J, Karovic S, et al. Enhanced detection of treatment effects on metastatic colorectal cancer with tumor burden growth rate evaluation. *Clin Cancer Res*. 2020;26(24):6464-6474.

6. Lubner MG, Stabo N, Lubner SJ, et al. Volumetric versus unidimensional measures of metastatic colorectal cancer in assessing disease response. *Clin Colorectal Cancer*. 2017;16(4):324-333.e1.
7. Wulff AM, Fabel M, Freitag-Wolf S, et al. Volumetric response classification in metastatic solid tumors on MSCT: initial results in a whole-body setting. *Eur J Radiol*. 2013;82(10):e567-e573.
8. Mozley PD, Bendtsen C, Zhao B, et al. Measurement of tumor volumes improves RECIST-based response assessments in advanced lung cancer. *Transl Oncol*. 2012;5(1):19-25.
9. Winter KS, Hofmann FO, Thierfelder KM, et al. Towards volumetric thresholds in RECIST 1.1: therapeutic response assessment in hepatic metastases. *Eur Radiol*. 2018;28(11):4839-4848.
10. Owen B, Gandara D, Kelly K, et al. CT volumetry and basic texture analysis as surrogate markers in advanced non-small-cell lung cancer. *Clin Lung Cancer*. 2020;21(3):225-231.
11. Xu J, Yin Y, Yang J, et al. Modified quantitative and volumetric response evaluation criteria for patients with hepatocellular carcinoma after transarterial chemoembolization. *Front Oncol*. 2023;13:957722.
12. Zhao B. Understanding sources of variation to improve the reproducibility of radiomics. *Front Oncol*. 2021;11:633176.
13. CT tumor volume change for advanced disease (CTV-AD). 2018. Accessed June 22, 2018. https://qibawiki.rsna.org/index.php/Archived_Versions#Advanced_disease
14. Dercle L, Connors DE, Tang Y, et al. PACT: an FNIH public-private partnership supporting sharing of clinical trial data for development of improved imaging biomarkers in oncology. *JCO Clin Cancer Inform*. 2018;2:1-12.
15. Yang H, Schwartz LH, Zhao B. A response assessment platform for development and validation of imaging biomarkers in oncology. *Tomography*. 2016;2(4):406-410.
16. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. 2012;30(9):1323-1341.
17. Welsh JL, Bodeker K, Fallon E, et al. Comparison of response evaluation criteria in solid tumors with volumetric measurements for estimation of tumor burden in pancreatic adenocarcinoma and hepatocellular carcinoma. *Am J Surg*. 2012;204(5):580-585.
18. Schiavon G, Ruggiero A, Schöffski P, et al. Tumor volume as an alternative response measurement for imatinib treated GIST patients. *PLoS One*. 2012;7(11):e48372.
19. Nishino M, Jackman DM, DiPiro PJ, et al. Revisiting the relationship between tumour volume and diameter in advanced NSCLC patients: an exercise to maximize the utility of each measure to assess response to therapy. *Clin Radiol*. 2014;69(8):841-848.
20. Gong AJ, Ruchalski K, Kim HJ, et al. RECIST 1.1 target lesion categorical response in metastatic renal cell carcinoma: a comparison of conventional versus volumetric assessment. *Radiol Imaging Cancer*. 2023;5(5):e220166.
21. Moertel CG, Hanley JA. The Effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976;38(1):388-394.
22. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer*. 1981;47(1):207-214.
23. Piessevaux H, Buyse M, Schlichting M, et al. Use of early tumor shrinkage to predict long-term outcome in metastatic colorectal cancer treated with cetuximab. *J Clin Oncol*. 2013;31(30):3764-3775.
24. Heinemann V, Stintzing S, Modest DP, et al. Early tumour shrinkage (ETS) and depth of response (DpR) in the treatment of patients with metastatic colorectal cancer (mCRC). *Eur J Cancer*. 2015;51(14):1927-1936.
25. Choi H, Charnsangavej C, Faria SC, et al. Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: proposal of new computed tomography response criteria. *J Clin Oncol*. 2007;25(13):1753-1759.

Radiology Advances, 2025, 2, 1–8
https://doi.org/10.1093/radadv/umaf001
Original Research