Published in final edited form as:
Clin Cancer Res. 2014 July 1; 20(13): 3560–3568. doi:10.1158/1078-0432.CCR-14-0245.

Variability in assessing treatment response: metastatic colorectal cancer as a paradigm

Binsheng Zhao[1], **Shing M Lee**[2], **Hyun-Ju Lee**[1], **Yongqiang Tan**[1], **Jing Qi**[1], **Thorsten Persigehl**[1], **David P Mozley**[3], and **Lawrence H Schwartz**[1]

[1]Department of Radiology, Columbia University Medical Center, 710 West 168th Street, NI-B26, New York, NY 10032

[2]Mailman School of Public Health, Columbia University, 722 W 268th St, Rm-645, New York, NY 10032

[3]Department of Imaging, Merck Research Laboratory, West Point, PA

Abstract

Purpose—The cut-off values currently used to categorize tumor response to therapy are neither biologically based nor tailored for measurement reproducibility with contemporary imaging modalities. Sources and magnitudes of discordance in response assessment in metastatic colorectal cancer (mCRC) are unknown.

Experimental Design—A subset of patients' CT images of chest, abdomen and pelvis were randomly chosen from a multi-center clinical trial evaluating IGF-1R targeted therapy in mCRC. Using Response Evaluation Criteria in Solid Tumors (RECIST), three radiologists selected target lesions and measured UNI (maximal diameter), BI (product of maximal diameter and maximal perpendicular diameter) and VOL (volume) on baseline and 6-week post-therapy scans in the following ways: (1) each radiologist independently selected and measured target lesions and (2) one radiologist's target lesions were blindly re-measured by the others. Variability in relative change of tumor measurements was analyzed using linear mixed effects models.

Results—Three radiologists independently selected 138, 101 and 146 metastatic target lesions in the liver, lungs, lymph nodes and other organs (e.g., peritoneal cavity) in 29 patients. Of 198 target lesions total, 33% were selected by all three, 28% by two, and 39% by one radiologist. With independent selection, the variability in relative change of tumor measurements was 11% (UNI), 19% (BI) and 22% (VOL), respectively. When measuring the same lesions, the corresponding numbers were 8%, 14% and 12%.

Corresponding author: Binsheng Zhao, Department of Radiology, Columbia University Medical Center, 710 West 168th Street, NI-B26, New York, NY 10032, bz2166@cumc.columbia.edu, Phone: 212-342-1393.

*Dr. Mozley was Merck's employee when he participated in this study

Current Demographics:
Hyun-Ju Lee, Department of Radiology, Seoul National University Hospital, 28 Yoengeon-dong, Jongno-gu, Seoul 110-744, Republic of Korea
Jing Qi, Mallinckrodt Institute of Radiology, 510 S. Kingshighway Blvd., St. Louis, MO 63110.
Thorsten Persigehl, Department of Radiology, University Hospital Cologne, Kerpener Str. 62, 50937 Cologne, Germany
David P Mozley, Department of Radiology, Weill Cornell Medical College, 520 East 70th Street, Starr Pavilion, 2nd Floor, New York, NY 10065

The authors of this manuscript have no conflicts of interest to disclose.

Conclusions—The relatively low variability in change of mCRC measurements suggests that response criteria could be modified to allow more accurate and sensitive CT assessment of anti-cancer therapy efficacy.

Keywords

RECIST; volume; variability; computed tomography; metastatic colorectal cancer

Introduction

Optimizing the use of quantitative imaging as a more accurate and early biomarker for tumor response is crucial to the development of new therapeutics and, further, for patient management. In drug discovery, the more promptly and precisely a go/no-go decision can be made, the better the chance for a successful trial. In clinical practice, response to a new therapy would ideally be determined with high accuracy as soon as possible after beginning the therapy to permit a change if in fact that therapy is ineffective.

Since the establishment of the Response Evaluation Criteria in Solid Tumors (RECIST) in 2000, unidimensional measurement has become the standard method for assessing tumor response to therapy in clinical trials and clinical practice (1,2). Based on the percentage change in the sum of the longest diameter (SLD) of all target lesions, RECIST reports tumor response to therapy using a 4-category system that was adopted from its predecessor, the World Health Organization (WHO) criteria (3,4). However, the response and progression cut-off values, e.g., 30% or more decrease in SLD, developed to be compatible with historical data, were neither biologically based nor tailored for measurement reproducibility with contemporary imaging modalities (5,6).

The limitations of RECIST are evident when evaluating tumor response to targeted therapies whose efficacy often does not correspond to rapid tumor shrinkage or shrinkage at all. For example, studies in metastatic colorectal cancer (mCRC) reported that patients, who received targeted therapies and showed minor early tumor shrinkage of about 10% or more, had better overall survival and progression-free-survival (7–10). Further, a phase II clinical study correlating early tumor radiographic change with *EGFR* mutation status in non-small cell lung cancer treated with Gefitinib found that the volumetric technique was significantly more sensitive and specific than the RECIST method at dichotomizing tumors into *EGFR* mutant and *EGFR* WT groups (11).

There are two principle sources of variability when applying RECIST to the assessment of mCRC. One is biologic heterogeneity of colorectal cancer (12): Not all lesions or metastatic foci grow at the same rate or respond to treatment in the same manner. Therefore the selection of target lesions must result in variability of measurements based on the sum of the longest diameters as a surrogate change in whole body tumor burden. Another source of variability is the performance of actual measurements. The purpose of this study was therefore to evaluate these two common sources of variability in the interpretation of early change of total tumor burden calculated unidimensionally, bidimensionally and volumetrically, using a subset of a contemporary mCRC multi-center clinical trial imaging data.

Materials and Methods

Patient image data

This study analyzed a randomly selected subset of the de-identified CT images taken from a completed multi-center Phase II/III clinical trial testing a therapy targeting the Insulin-like Growth Factor Receptor Type 1 (IGF-1R) in patients with mCRC (ClinicalTrials.gov identifier NCT00614393) (13). Case selection was based on the following criteria: the first 30 patients who had at least one measureable lesion at baseline per RECIST and had a follow-up scan at 6 weeks +/− 1 day.

The CRC clinical trial study used the standard CT imaging protocol in current clinical trial practice. The majority of the patients had a contrast-enhanced diagnostic chest scan reconstructed with a slice thickness of 5 mm and a sharper convolution kernel and a multi-phase contrast-enhanced diagnostic abdomen/pelvis scan reconstructed with a slice thickness of approximately 3 mm and a smoother convolution kernel.

Target lesion selection and measurement

Target lesions at each patient’s baseline study were selected per RECIST 1.0. After target lesion selection, each lesion was delineated on baseline and follow-up scan images using three 3-D lesion segmentation software developed for liver, lung and lymph node/peritoneal metastatic lesions, respectively (14–16). Computer-generated lesion contours were superimposed on the original images, reviewed and edited (if necessary) by radiologists in a side-by-side manner. The standard window/level settings of 1500/−500, 150/90 and 340/60 (in Hounsfield Unit) were used for reviewing and editing target lesions in the lungs, liver and lymph nodes (or peritoneal cavity), respectively. If there were multiple phases, portal vein phase images were used for measuring metastatic lesions in the abdomen and pelvis.

Based on the delineated contours, the longest axial-plane diameter (unidimensional measurement), its longest perpendicular diameter and the volume of a lesion were automatically calculated by computer program. The sums of unidimensional (SLD), bidimensional (SBI) and volumetric (SVOL) measurements of all target lesions in each patient, at baseline and 6-week follow-up scans and their relative changes at 6 weeks from baseline (SLD%, SBI% and SVOL%), were calculated for each radiologist’s reading.

Reading interpretation modes

Three radiologists (HJL, JQ and TP, with 11, 10 and 8 years’ experience of interpreting oncologic CT images, respectively) participated in the variability study. None were involved in study design or analysis.

Before beginning measurements, a training session was led by an expert radiologist (LHS) with more than 25 years’ experience interpreting CT images. Rules were reviewed, and relevant issues regarding target lesion selection were discussed. After the training session, there was no further communication between the readers and the trainer.

Two interpretation modes were evaluated in order to study biologic heterogeneity and tumor measurement variability.

Interpretation Mode #1 - Target lesion selection and measurement—

To study variability due to both lesion heterogeneity and measurement, each of the three radiologists independently selected target lesions at baseline and measured these target lesions at the baseline and 6-week follow-up scans.

Interpretation Mode #2 – Target lesion measurement only—

In the second experiment, only variability due to measurement was studied. Therefore, each radiologist independently measured a common set of pre-selected target lesions on the two scans of each patient. The common target lesions were adopted from one radiologist's (HJL) reading conducted in the first experiment. Each of the other two radiologists independently measured the common target lesions that had not been selected (measured) in their first experiment.

Statistical analysis

The baseline characteristics and measurements were summarized as counts and percentages for categorical variables and medians and inter-quartile ranges for continuous variables such as SLD, SBI and SVOL. Waterfall plots were created for baseline as well as relative change at 6 weeks. For the baseline plots, the variables were log-transformed.

To estimate the within-patient variability using the data from all three radiologists, linear mixed effects models with random intercept were fitted. The model accounted for the correlation among the three radiologists' measurements for each patient. The model based estimate for limits of agreement was +/−1.96 times the estimate of the within-patient standard deviation, i.e., the residual standard deviation. Analyses were performed using SAS 9.2 and STATA 12.0.

Results

Out of the 30 patients, one patient was excluded prior to statistical analysis because of the inconsistency among the three radiologists in including and measuring coalesced new lymph nodes on the follow-up scan. In the remaining 29 patients, twelve (41%) were 65 or older and eleven (38%) were female.

Interpretation mode #1 - target lesion selection and measurement

Table 1 shows that three radiologists selected 138, 101 and 146 metastatic lesions as target lesions in 29 patients, respectively. These target lesions were found in the liver (~50–60%), lungs (~30%), lymph nodes (~8–12%) and other organs (peritoneal cavity: abdominal wall=16:1) (~5–9%). In total, there were 198 different target lesions selected by at least one radiologist, among which 66 (33%) lesions were selected by all three radiologists, 55 (28%) by any two radiologists and 77 (39%) by only one radiologist. It is interesting to note that about 67% of lymph nodes were selected by only one radiologist, indicating that the most likely site for radiologists to pick up different metastatic lesions as target lesions is the lymph nodes. This could be explained by the widespread nature of lymph nodes.

The number of selected target lesions per patient and the radiologists' measurements of baseline SLD, SBI and SVOL are presented in Table 2. The median number of target lesions per patient selected by the three radiologists varied from 3 to 5 and the median total tumor burden measured at baseline varied from 10.4 cm to 16.3 cm (SLD), 26.1 cm^2 to 41.5 cm^2 (SBI) and 46.0 cm^3 to 72.2 cm^3 (SVOL). The maximal variations of the median total tumor burden (the maximal absolute difference among the three radiologists' median measurements divided by the average of their median measurements) were over 40% for all three measurement techniques.

The median SLD%, SBI% and SVOL% at 6-week post-therapy along with the inter-quartile ranges for radiologists 1, 2 and 3 are presented in Table 3. The median percentage change in total tumor burden measured by the three radiologists varied from −21% to −15% (SLD%), −35% to −25% (SBI%) and −45% to −41% (SVOL%). The within-patient variability in measuring SLD%, SBI% and SVOL% were +/−11%, +/−19% and +/−22%, respectively.

Figure 1 displays the waterfall plots of baseline SLD, SBI and SVOL and their relative changes at 6 weeks for each individual patient by radiologist. It shows that despite widely distributed measurements of total tumor burden at baseline (Fig. 1A–C), relative changes in total tumor burden among the three radiologists were within narrower ranges (Fig. 1D–F).

Looking at the distributions (at an increment of 5%) of the maximal differences of SLD%, SBI% and SVOL% measured among the three radiologists, Table 4 shows, for instance, that for 86% of the cases the variability in SLD% was within 15%, for 76% of the cases the variability in SBI% was within 20% and for 83% of the cases the variability in SVOL% was within 25%. With these cut-offs, about or above 80% of the cases' measurement variability fall into the ranges of 15% (SLD), 20% (SBI) and 25% (SVOL).

Interpretation mode #2 – target lesion measurement only

Tables 2 and 3 and Figure 2 present the distributions of baseline total tumor burdens and their relative changes at 6 weeks, when three radiologists measured the same target lesion set. The median values of radiologists' baseline measurements of SLD, SBI and SVOL varied from 15.1 cm to 16.3 cm, 39.1 cm^2 to 41.2 cm^2 and 59.7 cm^3 to 62.7 cm^3, respectively (Table 2). The maximal variations were about 5% for all three measurement techniques, much lower compared to the baseline measurements when considering the variable of target lesion selection.

The median SLD%, SBI% and SVOL% at 6 weeks along with the inter-quartile ranges for radiologists 1, 2 and 3 are presented in Table 3. The median percentage change in total tumor burden measured by the three radiologists varied from −24% to −20% (SLD%), −39% to −35% (SBI%) and −46% to −41% (SVOL%). The within-patient variability in measuring SLD%, SBI% and SVOL% were +/−8%, +/−14% and +/−12%, lower than if also considering the variable of target lesion selection.

Figure 2, the waterfall plots of the baseline SLD, SBI and SVOL (Fig. 2A–C) and their relative changes at 6 weeks (Fig. 2D–F) for each individual patient by radiologist, shows once again that variability in the measurement of total tumor burden was greatly reduced when radiologists measured the same set of target lesions compared to when they selected target lesions individually. However, it is interesting to note that the distribution patterns and variation ranges of the relative changes in total tumor burden were similar for both interpretation modes.

Table 4 shows that for all of the cases the variability in SLD% was within 15%, for 86% of the cases the variability in SBI% was within 20% and for 90% of the cases the variability in SVOL% was within 25%.

Discussion

In addition to the objective response rate, many other metrics used in oncology clinical trials as surrogate endpoints (e.g., progression free survival, time-to-progression) are also based on tumor change detected from longitudinal CT examinations. Understanding variability in the interpretation of tumor change over time is therefore essential, as true tumor biological change can be reliably determined only when it is above the magnitude of the measurement variability. Despite the widespread use of linear measurements (mainly unidimensional) in clinical trials and clinical care and the potential adoption of the volumetric technique, a more accurate and sensitive method for quantifying tumor change, the magnitude of variability of these measurements is not well known.

An increasing body of literature addresses the issue of measurement variability in the context of therapy response assessment in oncology (17–22). In those studies, radiologists were asked to independently measure/re-measure a pre-selected set of lesions at a single scan time-point. Though these studies provide the foundation for understanding measurement variability, they do not consider several other variables that may impact the estimation of tumor change. For instance, in order to evaluate response to therapy in metastatic cancer, the lesions in multiple organs must be monitored and, if not all lesions are to be measured, target lesions must be selected. How does the selection of target lesions affect the estimation of change in total tumor burden? Furthermore, over the course of therapy, both lesions and their surrounding organs/tissues can change. How do such changes affect the consistency of tumor measurements over time?

The unique design of our study differentiates it from previously published variability studies: 1) in addition to measurement-induced variability, we considered the effect of target lesion selection, 2) we investigated variability based on patient level (i.e., total tumor burden of all involving sites) rather than on the lesion level in a single site as the majority of studies did, 3) we explored the variability in relative change of total tumor burden from baseline to the early follow-up scan at 6 weeks (i.e., objective response rate used in response assessment), and 4) we studied variability in three measurement techniques, i.e., uni, bi and volume.

Our study found that when radiologists independently selected target lesions in mCRC patients, only 1/3 of the target lesions were selected by all three radiologists and a little more than 1/4 were selected by any two radiologists. More than 50% of the target lesions were picked up from the liver and about 30% were from the lungs. The remaining 20% were found in the lymph nodes (~10%) and peritoneal cavity (~10%).

Baseline total tumor burden has shown to be predictive for patient survival (23,24). First, we looked at the effect of target lesion selection on the estimation of total tumor burden at baseline. It was obvious that fewer target lesions often resulted in a smaller total tumor burden. Compared to measuring the same lesion set, the selection of target lesions drastically increased the variations of different radiologists' baseline measurements; the maximal variations of the median total tumor burden increased from about 5% to more than 40% for all three measurement techniques. However, target lesion selection only increased the within-patient variability in measuring early percentage changes in total tumor burden from 8% to 11% (SLD), 14% to 19% (SBI) and 12% to 22% (SVOL). Our findings suggest that target lesion selection can have a large effect on baseline measurement but a relatively small effect on measuring tumor change.

For each interpretation mode, we also studied intra-reader variability by asking one of the participating radiologists to repeat the measurements in a separate session. We found that for each measurement technique, intra-reader and inter-reader variability in relative change were of a similar magnitude when measuring the same lesion set. However, intra-reader variability was lower than inter-reader variability when involving target lesion selection. Due to the considerably large volume of data already presented, we did not report this part of the study.

The levels of measurement variability obtained in this study could be lower than that in the "real world". We did not consider possible human errors such as mismatching of lesions on longitudinal scans of a patient, as we wanted to study the variability caused only by the selection and measurement of target lesions. Moreover, our side-by-side reading manner and the reader training session held prior to the beginning of the study all helped to reduce variability. Other techniques may also reduce variance, such as well-controlled imaging acquisition guidelines as proposed by the Quantitative Imaging Biomarker Alliance (QIBA) (25), standardized and optimized thin-section CT imaging acquisition protocols (26, 27) and reducing human error through computer-aided quality assurance programs.

Based on our findings of within-patient variability, a conservative suggestion for the cutoff values to identify tumor changes in mCRC would be +/−15% for SLD, +/−25% for SBI and +/−30% for SVOL. These values are lower than the corresponding response cutoff values of −30%, −50% and −65% as proposed by RECIST/WHO guidelines. This indicates that we may be able to detect tumor change at a lower magnitude and/or earlier with the variability-based cut-off values. Furthermore, the spherical relationships among the cut-off values of unidimensional, bidimensional and volumetric measurements do not seem to exist. Given large measurement values, the volumetric technique seems to yield lower variability than the unidimensional technique. Correlations with clinical outcomes and other biomarkers are required, however, before drawing a conclusion as to the superiority of any one of these techniques, including volumetric measurements.

Limitations of this study include the small number of patients and participating radiologists. However, in many clinical trials, especially phase II, the number of study participants is similar to our study population. Also, many clinical trials rely on a small group of radiologists to serve as the reference radiologists, again, similar in number to our study.

Furthermore, tumor diameters and volume were measured using the algorithms developed from a single research laboratory. As we know, different segmentation algorithms may generate different segmentation results and thus different diameter and volume measurements. However, since we used the same algorithm to segment the same type of tumors on baseline and follow up scans, algorithm-intrinsic biases should be reduced to a minimum. It is worth mentioning that the evaluation and validation of lung nodule/lesion segmentation algorithms developed by both industry and institution are currently underway (25, 28). Additionally, commercial vendors of both imaging hardware and workstations are beginning to offer tumor delineation, volume measurement and rendering software.

The imaging techniques and the ways the image data were collected and metastatic lesions selected and measured in this study represent the current practice of multi-center clinical trials not only in mCRC but also other metastatic cancers. Our finding of relatively low variability associated with tumor change measurement will allow for potentially lowering response cut-off values of current criteria and establishing a new volumetric response method for more sensitive CT assessment of the efficacy of anti-cancer therapies. More sensitive response assessments would enhance patient care by sparing patients continued exposure to the toxicity of futile treatments and giving them earlier access to alternative treatments.

Acknowledgments

The research was supported in part by a sub-award of RSNA Quantitative Imaging Biomarker Alliance (QIBA) to Dr. Binsheng Zhao thru the NIH grant (Award No.: NHLBI-PB-EB-2010-159-JKS) awarded to Dr. Daniel C Sullivan and in part by the NIH grant U01 CA140207 awarded to Drs Lawrence H Schwartz and Binsheng Zhao.

References

1. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate response to treatment in solid tumors. J Natl Cancer Inst. 2000; 92:205–216. [PubMed: 10655437]

2. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1. 1). Eur J Cancer. 2009; 45:228–247. [PubMed: 19097774]

3. World Health Organization Offset Publication No 48. Geneva (Switzerland): 1979. WHO handbook for reporting results of cancer treatment.

4. Miller AB, Hogestraeten B, Staquet M, Winkler A. Reporting results of cancer treatment. Cancer. 1981; 47:207–214. [PubMed: 7459811]

5. Morel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. Cancer. 1976; 38:388–94. [PubMed: 947531]

6. Warr D, Mckinney S, Tannock I. Influence of measurement error on assessment of response to anticancer chemotherapy: proposal for new criteria of tumor response. J Clin Oncol. 1984; 2:1040–1046. [PubMed: 6206206]

7. De Roock W, Piessevaux H, De Schutter J, Janssens M, De Hertogh G, Personeni N, et al. KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. Ann Oncol. 2008; 19:508–15. [PubMed: 17998284]

8. Piessevaux H, Buyse M, De Roock W, Prenen H, Schlichting M, Van Cutsem E, et al. Radiological tumor size decrease at week 6 is a potent predictor of outcome in chemorefractory metastatic colorectal cancer treated with cetuximab (BOND trial). Ann Oncol. 2009; 20:1375–1382. [PubMed: 19465422]

9. Suzuki C, Blomqvist L, Sundin A, Jacobsson H, Bystrom P, Berglund A, et al. The Initial change in tumor size predicts response and survival in patients with metastatic colorectal cancer treated with combination chemotherapy. Ann Oncol. 2012; 23:948–54. [PubMed: 21832285]

10. Piesseraux H, Buyse M, Schlichting M, Van Cutsem E, Bokemeyer C, Heeger S, et al. Use of early tumor shrinkage to predict long-term outcome in metastatic colorectal cancer treated with cetuximab. J Clin Oncol. 2013; 31:3764–75. [PubMed: 24043732]

11. Zhao B, Oxnard GR, Moskowitz CS, Kris MG, Pao W, Guo P, et al. A pilot study of volume measurement as a method of tumor response evaluation to aid biomarker development. Clin Cancer Res. 2010; 16:4647–53. [PubMed: 20534736]

12. De Roock W, De Vriendt V, Normanno N, Ciardiello F, Teipar S. KRAS, BRAF, PIK3CA, and PTEN mutations: implications for targeted therapies in metastatic colorectal cancer. The Lancent Oncology. 2011; 12:594–603.

13. Atzori F, Tabernero J, Cervantes A, Prudkin L, Andrew J, Rodríguez-Braun E, et al. A Phase I Pharmacokinetic and Pharmacodynamic Study of Dalotuzumab (MK-0646), an Anti-Insulin-like Growth Factor-1Receptor Monoclonal Antibody, in Patients with Advanced Solid Tumors. Clin Cancer Res. 2011; 17:6304–12. [PubMed: 21810918]

14. Guo, X.; Zhao, B.; Schwartz, LH. Columbia University Invention Report IR #2906. Computer-aided tumor segmentation using local region-based active contours.

15. Tan Y, Schwartz LH, Zhao B. Segmentation of lung tumors on CT scans using watershed and active contours. Med Phys. 2013 Apr.40(4):043502. [PubMed: 23556926]

16. Tan, Y.; Zhao, B.; Schwartz, LH. Columbia University Invention Report CU #12256. Segmentation of lymph nodes by sphere subdivision and active contours.

17. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozafiak JA, et al. Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol. 1996; 167(4):851–4. [PubMed: 8819370]

18. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. J Clin Oncol. 2003; 21(13):2574–82. [PubMed: 12829678]

19. Punnen S, Haider MA, Lockwood G, Moulding F, O'Malley ME, Jewett MA. Variability in size measurement of renal masses smaller than 4 cm on computerized tomography. J Urol. 2006; 176(6):2386–90. [PubMed: 17085106]

20. Zhao B, James LP, Moskowitz CS, Guo P, Ginsberg MS, Lefkowitz RA, et al. Evaluating variability in tumor measurements from same-day repeat CT scans in patients with non-small cell lung cancer. Radiology. 2009; 252(1):263–72. [PubMed: 19561260]

21. Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, et al. Variability of lung tumor measurements on repeat CT scans taken within 15 minutes: implications for care and clinical research. J Clin Oncol. 2011; 29(23):3114–9. [PubMed: 21730273]

22. McErlean A, Panicek DM, Zabor EC, Moskowitz CS, Bitar R, Motzer RJ, et al. Intra- and Interobserver Variability in CT Measurements in Oncology. Radiology. 2013; 269:451–9. [PubMed: 23824993]

23. Pass HI, Temeck BK, Kranda K, Steinberg SM, Feuerstein IR. Preoperative tumor volume is associated with outcome in malignant pleural mesothelioma. J Thorac Cardiovasc Surg. 1998; 115(2):310–8. [PubMed: 9475525]

24. Liu F, Zhao B, Krug LM, Ishill NM, Lim RC, Guo P, et al. Assessment of therapy responses and prediction of survival in malignant pleural mesothelioma through computer-aided volumetric measurement on CT scans. J Thorac Oncol. 2010 Jun; 5(6):879–84. [PubMed: 20421814]

25. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC, Aerts HJ, Bendriem B, et al. Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. Radiology. 2011; 259:875–84. [PubMed: 21325035]

26. Tan Y, Guo P, Mann H, Marley SE, Juanita Scott ML, Schwartz LH, et al. Assessing the effect of computed tomographic (CT) slice thickness on unidimensional (1D), bidimensional (2D) and volumetric measurements of solid tumors. Cancer Imaging. 2012; 12:497–505. [PubMed: 23113962]

27. Zhao B, Tan Y, Bell DJ, Marley SE, Guo P, Mann H, et al. Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors in CT scans reconstructed at different slice intervals. Eur J Radiol. 2013; 82(6):959–68. [PubMed: 23489982]

28. It is worth mentioning that the evaluation and validation of lung nodule/lesion segmentation algorithms developed by both industry and institution are currently underway (25, 28). Additionally, commercial vendors of both imaging hardware and workstations are beginning to offer tumor delineation, volume measurement and rendering software.

Translational Relevance

Targeted cancer therapy results in different patterns of response in solid tumors, including smaller magnitude of size change than cytotoxic therapy. Studies in colorectal cancer demonstrate that tumor "minor" reduction is a predictor of Overall Survival and Progression-Free-Survival. Investigators are finding that more accurate classifications of response can be derived on the basis of biology than the historical response criteria. Such a biomarker or metric of response can be universally used for multiple tumors and targeted therapies. However, a biomarker must be both biologically meaningful and reproducibly measurable by imaging. This study explores the latter, by assessing variability in measuring tumor change in a metastatic setting and also considering heterogeneity of target lesion selection, a previously unstudied factor. Our finding of low measurement variability allows for improving historical response criteria by lowering the response cut-off and/or introducing the volumetric technique as a more sensitive CT assessment of treatment efficacy.

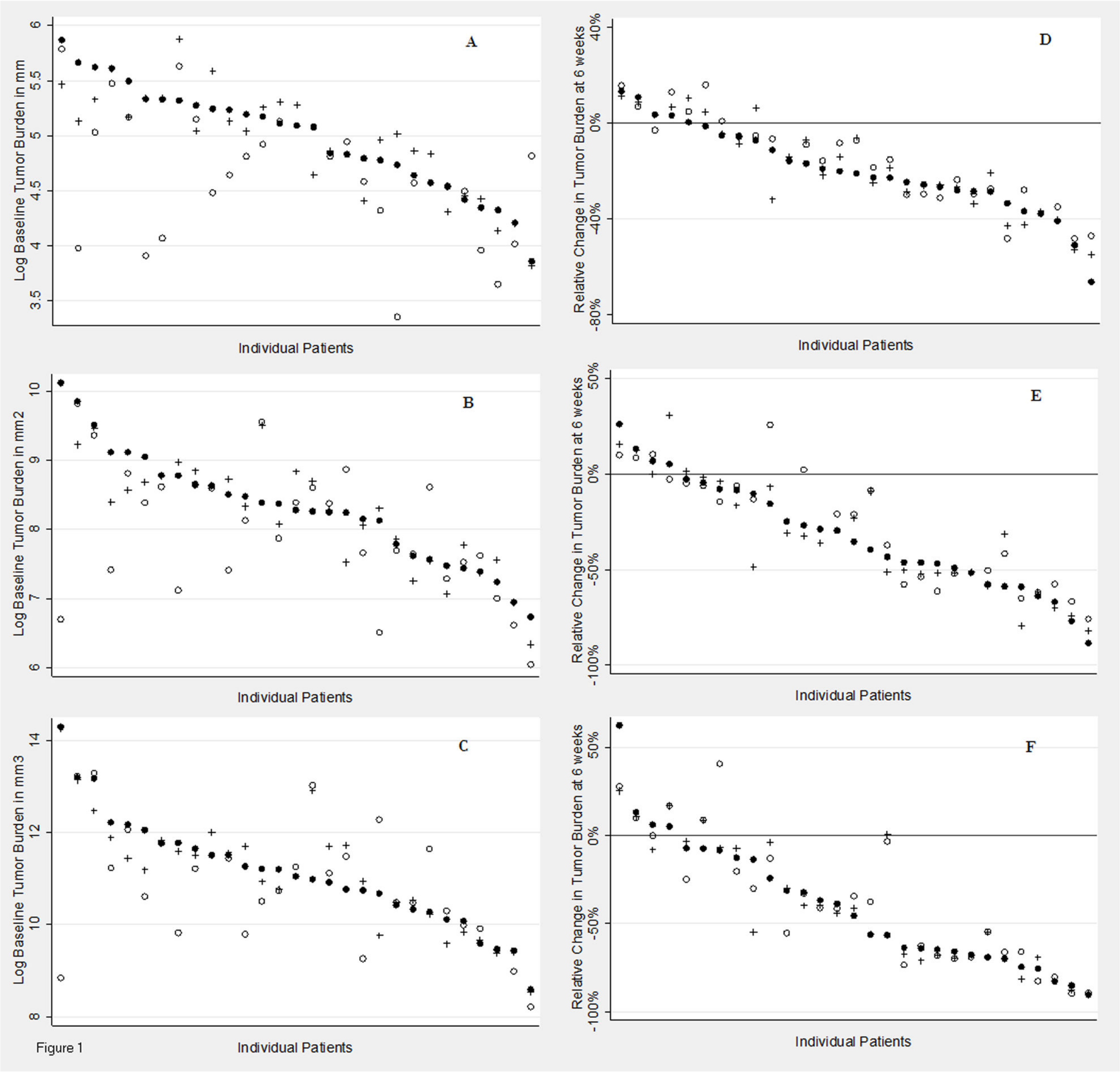

Figure 1.

Waterfall plots of the baseline SLD (A), SBI (B) and SVOL (C) and their relative changes at 6-week (D–F). Patient’s order in the waterfall plots was determined based on radiologist 1’s baseline measurements (solid circle). The solid circle, hollow circle and + sign represent radiologists 1, 2 and 3, respectively.

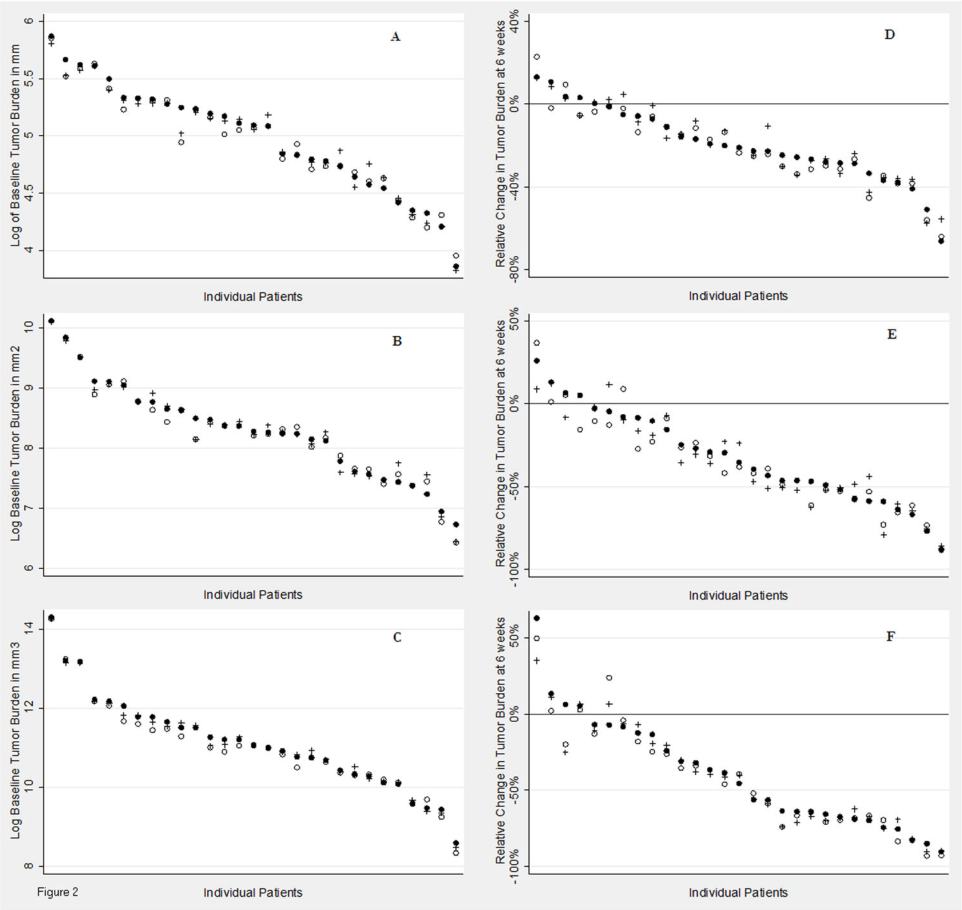

Figure 2.

Waterfall plots of the baseline SLD (A), SBI (B) and SVOL (C) and their relative changes at 6-week (D–F). Patient's order in the waterfall plots was determined based on radiologist 1's baseline measurements (solid circle). The solid circle, hollow circle and + sign represent radiologists 1, 2 and 3, respectively.

Table 1

Distributions of the target lesions selected by radiologists in different sites

	Target lesion distributions (29 patients)				
Individually selected by Radiologist	**Liver**	**Lung**	**Lymph node**	**Others**	**Total lesion selected**
1	70 (50.7%)	40 (29.0%)	16 (11.6%)	12 (8.7%)	138
2	59 (58.4%)	29 (28.7%)	8 (7.9%)	5 (5.0%)	101
3	76 (52.1%)	44 (30.1%)	14 (9.6%)	12 (8.2%)	146

Commonly selected (no repetition) by Radiologists	Liver	Lung	Lymph node	Others	
All three	38 (38%)	20 (35.1%)	6 (25%)	2 (11.8%)	66 (33.3%)
Any two	29 (29%)	16 (28.1%)	2 (8.3%)	8 (47.1%)	55 (27.8%)
Only one	33 (33%)	21 (36.8%)	16 (66.7%)	7 (41.2%)	77 (38.9%)
Total lesions	100	57	24	17	198

Table 2

Distribution of baseline total tumor burden per patient

Interpretation Mode	Radiologist	Lesion Number		SLD (cm)		SBI (cm^2)		SVOL (cm^3)	
		Median	IQR*	Median	IQR*	Median	IQR	Median	IQR
# 1 Target lesion selection and measurement	1	5	(3, 5)	16.3	(10.4, 20.5)	39.5	(20.3, 64.6)	62.3	(30.8, 129.3)
	2	3	(3, 4)	10.4	(7.6, 16.0)	26.1	(16.6, 54.8)	46.0	(20.2, 99.5)
	3	5	(4, 5)	15.5	(10.4, 20.3)	41.5	(19.2, 64.9)	72.2	(27.4, 124.3)
# 2 Target lesion Measurement Only	1	5	(3, 5)	16.3	(10.4, 20.5)	39.5	(20.3, 64.6)	62.3	(30.8, 129.3)
	2	5	(3, 5)	15.1	(10.8, 20.2)	41.2	(21.3, 56.5)	59.7	(30.2, 99.3)
	3	5	(3, 5)	15.7	(11.6, 19.8)	39.1	(20.0, 65.7)	62.7	(32.5, 113.2)

[*] IQR: the inter-quartile range.

In mode #2, Radiologists 2 and 3 measured the same target lesion set determined by radiologist 1 during the first experiment

Table 3

Relative changes in total tumor burden at 6-week post-therapy

Interpretation Mode	Radiologist	SLD (%)		SBI (%)		SVOL (%)	
		Median	IQR	Median	IQR	Median	IQR
# 1 Target lesion selection and measurement	1	−21	(−28, −6)	−35	(−52, −8)	−45	(−68, −12)
	2	−15	(−30, −5)	−25	(−54, −6)	−41	(−68, −13)
	3	−21	(−32, −5)	−33	(−52, −7)	−44	(−69, −7)
Within-patient variability		+/−11%		+/−19%		+/−22%	
# 2 Target lesion measurement Only	1	−21	(−28, −6)	−35	(−52, −8)	−45	(−68, −12)
	2	−24	(−32, −6)	−39	(−53, −16)	−46	(−70, −20)
	3	−20	(−34, −6)	−36	(−52, −10)	−41	(−69, −19)
Within-patient variability		+/−8%		+/−14%		+/−12%	

Table 4

Maximal absolute differences of SLD%, SBI% and SVOL% among three radiologists

Interpretation Mode		SLD%				SBI%				SVOL%		
			Frequency	%	Cumulative %		Frequency	%	Cumulative %	Frequency	%	Cumulative %
	#1: Target lesion selection & measurement	≤5%	8	28	28	≤5%	5	17	17	9	31	31
		5–10%	12	41	69	5–10%	6	21	38	4	14	45
		10–15%	5	17	86	10–15%	10	34	72	6	21	66
		15–20%	3	10	97	15–20%	1	3	76	3	10	76
		>20%	1	3	100.0	20–25%	1	3	79	2	7	83
						25–30%	1	3	83	1	3	86
						>30%	5	17	100.0	4	14	100.0
	# 2: Target lesion measurement only	≤5%	12	41	41	≤5%	5	17	17	10	34	34
		5–10%	12	41	83	5–10%	9	31	48	11	38	72
		10–15%	5	17	100.0	10–15%	6	21	69	5	17	90
						15–20%	5	17	86	0	0	90
						20–25%	3	10	97	0	0	90
						25–30%	1	3	100.0	1	3	93
										2	7	100.0