Published in final edited form as:
Cancer. 2014 March 1; 120(5): 711–721. doi:10.1002/cncr.28493.

Intra- and Interobserver Variability in CT Size and Attenuation Measurements in Renal Cell Carcinoma Patients on Anti-Angiogenic Therapy: Implications for Alternative Response Criteria

Katherine Krajewski, MD[1,*], **Mizuki Nishino, MD**[1], **Yoko Franchetti, PhD**[2], **Nikhil Ramaiya, MD**[1], **Annick Van den Abbeele, MD**[1], and **Toni Choueiri, MD**[3]

[1]Department of Imaging, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA

[2]Department of Biostatistics, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA

[3]Kidney Cancer Center, Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, MA

Abstract

Background—Alternative response criteria have been proposed in patients with metastatic Renal Cell Carcinoma (mRCC) on Vascular Endothelial Growth Factor (VEGF)-targeted therapy, including 10% tumor shrinkage as an indicator of response/outcome. However, intraobserver and interobserver measurement variability have not been defined in this setting. We aim to determine intra- and interobserver agreement of Computed Tomography (CT) size and attenuation measurements, to establish reproducible response indicators.

Methods—Seventy-one mRCC patients with 179 target lesions were enrolled in Phase II and III trials of VEGF-targeted therapies and retrospectively studied with institutional review board approval. Two radiologists independently measured long axis diameter and mean attenuation of targets on baseline and follow-up CT. Concordance correlation coefficients (CCCs) and Bland-Altman plots were used to assess intra- and interobserver agreement.

Results—High CCCs (0.8602–0.9984) were observed in all types of measurements. The 95% limits of agreement for percent change of the sum longest diameter was (−7.30%, 7.86%) for intraobserver variability, indicating 10% tumor shrinkage represents true change in tumor size when measured by one observer. The 95% limits of interobserver variability were (−16.3%, 15.4%). In multivariate analysis, liver location significantly contributed to interobserver variability (p=0.048). The 95% limits of intraobserver agreement for percent change in CT attenuation were (−18.34%, 16.7%).

[*]Corresponding author: 450 Brookline Ave, Boston, MA 02115, kmkrajewski@partners.org, Phone: 617-582-8088/Fax: 617-632-3581.

Disclosure of Potential Conflicts of Interest:
KMK: research grant from General Electric and Association of University Radiologists. TKC: advisory boards for Pfizer/Aveo/GlaxoSmithKline/Novartis; research support from Pfizer. No potential conflicts of interest were disclosed by other authors.

Conclusion—In mRCC patients treated with VEGF-inhibitors, 10% tumor shrinkage is a reproducible radiologic response indicator when baseline and follow-up studies are measured by a single radiologist. Lesion location contributes significantly to measurement variability and should be considered when selecting target lesions.

Keywords

Renal Cell Carcinoma; Computed Tomography; RECIST; Tumor Shrinkage; CT Attenuation; Intraobserver Variability; Interobserver Variability

INTRODUCTION

Kidney cancer is the tenth leading cause of cancer death in men in the United States, with 13,570 estimated deaths in Americans in 2012 and 64,770 estimated new cases (1). At diagnosis, approximately 20–30% of patients with RCC demonstrate metastatic disease and 25–50% of patients with locoregional disease at diagnosis eventually develop metastases (2, 3). Therefore, systemic therapy is indicated in a significant number of patients, yet RCC is resistant to conventional chemotherapy. Vascular Endothelial Growth Factor (VEGF)-targeted treatments have become standard in metastatic renal cell carcinoma (mRCC; 4), a setting in which anti-tumor activity is evidenced by prolonged progression-free survival in treated patients, in spite of different rates of tumor shrinkage (5–8).

Oncologists rely on imaging to assess changes in tumor size, as detected by computed tomography (CT) scans, for evidence of response to therapy or disease progression in determining when to continue a therapy or consider alternative treatment. Response Evaluation Criteria in Solid Tumors (RECIST) is the widely accepted methodology to determine objective response, based on the sum of the longest unidimensional diameters (SLD) of target lesions (9). However, less than half of RCC patients treated with VEGF-targeted agents achieve response by RECIST, which requires 30% decrease in SLD of target lesions, even though prolonged time on therapy has been noticed in patients whose tumor shrinkage is less than 30% (5–8). For example, in the recent pazopanib versus sunitinib study, the response rate was 31% for pazopanib and 25% for sunitinib (10). RECIST-based response assessment may not be optimal to accurately evaluate anti-tumor activity in this setting, and alternative criteria for response are needed.

Emerging data support alternative imaging criteria to define “response” to VEGF-targeted therapies in mRCC, (11–17). Recently, a 10% tumor shrinkage threshold has been advocated as an indicator of response (11, 15). In our recent study of 70 mRCC patients treated with VEGF- inhibitors, 10% tumor shrinkage at first follow-up was a significant predictor of time to treatment failure (TTF) and overall survival (OS) (15). Other criteria, such as Choi, utilizing 10% decrease in size or 15% decrease in CT density for response, have also been applied to mRCC patients (11, 16, 17). While data supports use of a smaller tumor size change cut-off to define responders (10% decrease as opposed to 30% in RECIST), intra and interobserver measurement variability in this population has not been established. It is unknown whether 10% tumor shrinkage is indicative of tumor size change or within measurement error. To apply a new tumor shrinkage threshold indicative of response, it is

important to assess measurement reproducibility in mRCC patients treated in prospective trials of VEGF-targeted agents, to ensure that the response threshold is robust and reproducible.

The purpose of this study is to define intraobserver and interobserver variability of tumor size and attenuation measurements in mRCC patients treated with anti-angiogenic therapy. We hypothesize that 10% tumor shrinkage (i.e. 10% decrease in the SLD of target lesions) represents true change in tumor burden. Our goal is to advance the application of widely available CT technology, to define the best, most reliable, indicator of treatment response.

METHODS

Patients and treatment

The study sample consists of patients with metastatic renal cell carcinoma enrolled in six recent phase II and III multicenter, open-label studies of VEGF-targeted agents (tivozanib, pazopanib, foretinib, sorafenib, vatalanib, and sunitinib). Patients enrolled in these trials who had been included in our previous study cohort were excluded from this validation cohort (15). All patients had histologically confirmed metastatic RCC. Patients were imaged and treated at a single institution, at standard doses of study drug according to the assigned protocol until they experienced disease progression, severe toxicity, or withdrew consent. Compliance was checked after each cycle with a treatment diary. Patients were part of IRB-approved protocols for mRCC at the institution where baseline and follow-up clinical data was prospectively collected.

Imaging and image analysis

Patients eligible for analysis included those with target lesions by RECIST who underwent non-contrast-enhanced or contrast-enhanced CT of the chest, abdomen, and pelvis prior to and after VEGF-targeted therapy initiation, with pre- and post-therapy scans at the same institution. The routine oncology protocol was employed using a 64-row MDCT scanner (Aquilion 64; Toshiba America Medical Systems, CA, USA) or a 4-row MDCT scanner (Volume Zoom; Siemens Medical Solutions, Forchheim, Germany). Imaging parameters were as follows: (1) 64-row MDCT scanner: 0.5 mm collimation, 120kVp, 100–500 mA using dose modulation with noise index of 12.0 HU, 0.5s gantry rotation time and a table speed of 53 mm per rotation; (2) 4-row MDCT scanner: 2.5 mm collimation, 120kVp, 165mAs, 0.5s gantry rotation time and a table speed of 11.5 mm per rotation. All patients were scanned from cranial to caudal direction from the clavicles to the adrenal glands on supine position. Patients were instructed to maintain suspended inspiration during the CT acquisition. During the study, 75–100 mL of iopromid (Ultravist 300, 300 mg iodine/mL; Bayer HealthCare Pharmaceuticals Inc. Wayne, NJ, based on eGFR) was injected intravenously with an automated injector (Medrad, Pittsburgh, PA) at a rate of 2–3 mL/sec in patients with adequate estimated glomerular filtration rate (eGFR) and no known allergy. Scan delay for the chest was 30 seconds for 64-row MDCT and 40 seconds for 4-row MDCT, and 70 seconds for the abdomen. Axial images (5mm thickness and 5 mm interval) were reconstructed using standard and lung algorithms. Coronal reformatted images (5mm thickness and 5 mm interval) were also reconstructed for 64-row MDCT. Images were

reviewed and measured on Picture Archiving and Communication System (Centricity, General Electric, Milwaukee, WI).

Up to three target lesions in each patient were selected on the baseline CT by a single oncoradiologist (KK, Radiologist 1), blinded to follow-up CT changes and patient outcomes. Target lesions were selected and measured according to RECIST 1.0 (9), in keeping with the use of long axis diameter measurements in the evaluation of % tumor shrinkage in alternative response criteria previously described (11–16). The location of the target lesion was recorded. At baseline and first follow-up, the longest axial axis of each target was recorded to the nearest millimeter by two board-certified radiologists with cancer imaging expertise (KK, Radiologist 1 and MN, Radiologist 2) as described previously (15). In addition, the average CT attenuation coefficient of target lesions on contrast-enhanced studies were measured on the most representative axial image and measured in Hounsfield units (HU) by drawing a freehand region of interest around the perimeter of the target, as large as possible to cover the lesion without extending outside of the lesion, as previously described (12, 15–18). The average HU within the region was calculated on the PACS workstation and recorded.

During each session of measurements, a radiologist first performed baseline followed by follow-up measurements for each patient, in succession, referring to the baseline measurements. The SLD and mean attenuation of the targets at baseline and follow-up were also recorded. Radiologist 1 performed two sessions of measurements that were two months apart, in random patient order, to assess intraobserver variation. Radiologist 2 performed each measurement once, also in a random order, to assess interobserver variation compared to Radiologist 1.

For each patient, percent change in SLD and mean tumor attenuation per patient were calculated from measurements in each session (two sets for Radiologist 1, one for Radiologist 2). The percent changes in longest diameter and CT attenuation of each lesion were also calculated.

Statistical Analysis

Intra- and interobserver variability were assessed using: (a) concordance correlation coefficients (CCCs), (b) mean relative difference (%) with standard deviation (SD), and (c) 95% limits of agreement (the mean relative change ± 1.96 standard deviations of the difference) for the longest diameter and average CT attenuation assuming that the distributions of size and attenuation are normal. CCCs are products of a measure of precision (defined by Pearson's correlation) and a measure of accuracy (defined by a bias correction factor) where CCC value 1 indicates perfect agreement and −1 indicates perfect reversed agreement (19). The mean relative difference (%) between the two measurements is defined as $100*[M_1-M_2]/M_1$ (M_1 = measurement 1, M_2 = measurement 2) of the difference among all tumors.

Agreement in the two measurements was shown visually using Bland-Altman plots with 95% limits of agreement and the average relative difference, computing the mean relative difference (%) between the two measurements, plotted against the first measurement of

Radiologist 1 (20–22). Two measurements of Radiologist 1 were used to assess intraobserver variability. The first measurement of Radiologist 1 and the measurement by Radiologist 2 were used to evaluate interobserver variability.

Kappa analysis was performed to assess agreement between binary responses (>10% decrease and ≤ 10% decrease), to evaluate the impact of measurement variability on response assessment according to 10% tumor shrinkage criteria. Agreement between two assessments was categorized as poor ($\kappa < 0$), slight ($\kappa = 0$–0.20), fair ($\kappa = 0.21$–0.40), moderate ($\kappa = 0.41$–0.60), substantial ($\kappa = 0.61$–0.80), and almost perfect ($\kappa > 0.80$) (23).

The influence of lesion location on measurement variability of size and attenuation was assessed using multivariate linear regression models assuming a constant bias correction factor that comprise CCCs. Analyses were performed using MedCalc (MedCalc Software bvba, Ostend, Belgium) and SAS 9.2 (SAS Institute Inc., Cary, NC).

RESULTS

Size Measurements

Baseline size was measured in 71 patients with a total 179 target lesions. Follow-up size measurements were performed in 69 patients with 173 targets. Follow-up imaging was not available for review in two patients (one patient had progressive disease according to the radiology report but archived images from 2002 were not available, while another patient transferred care to another institution out of state). In a third patient, a single peritoneal target lesion was obscured by ascites and not identified on follow-up.

Sum Long Axis Diameter Measurements per Patient—Table 1 demonstrates the summary of intraobserver and interobserver variability in the SLD measurements per patient at baseline and percent change in SLD at first follow-up CT. Figure 1 shows the Bland-Altman plots for intra and interobserver variability of the measurements, with the mean percent difference and 95% limits of agreement of the two measures.

For intraobserver variability comparing two measurements by Radiologist 1, CCC was high for both baseline measurements and the percent change at the follow-up (0.9984, 0.9747, respectively). The 95% limits of agreement of the two measures (%) was (−7.90%, 7.17%) for baseline measurements, and (−7.30%, 7.86%) for the percent change on follow-up. Ten percent tumor shrinkage is beyond the 95% limits of agreement and therefore can be considered true change rather than measurement error, when baseline and follow-up measurements were performed by a single radiologist.

For interobserver variability comparing measurements by Radiologist 1 versus Radiologist 2, high CCC were also observed for both baseline measurements and percent change at first follow-up (0.9875, 0.8969, respectively). The range of 95% limits of agreement of the two independent measures was, however, approximately twice wider than the intraobserver assessment, and were (−17.0%, 16.6%) for baseline measurements and (−16.3%, 15.4%) for percent change at follow-up. The 10% tumor shrinkage threshold is within the range of 95% limits of agreement of interobserver variability.

Impact of measurement variability on response assessment using the 10% tumor shrinkage threshold was further investigated, demonstrated in a scatterplot (Fig. 2). Using 10% tumor shrinkage as the indicator of response, response assessment by two measurements Radiologist 1 had almost perfect agreement (κ=0.826). Response assessment by measurements of Radiologists 1 and 2 had substantial agreement (κ=0.682).

In a preliminary analysis to explore the effect of IV contrast-enhanced CT versus non-contrast CT on measurement variability, we performed CCC analysis for baseline SLD measurements in 54 patients with contrast versus 17 patients without contrast. For the 54 patients with contrast, intraobserver CCC was 0.9984 (0.9973–0.9991) and interobserver CCC was 0.9863 (0.9766–0.9920), while in 17 patients without contrast, intraobserver CCC was 0.9983 (0.9955–0.9994) and interobserver CCC was 0.9900 (0.9737–0.9963).

Long Axis Diameter Measurements per Lesion—Intra- and interobserver variability were studied in individual lesions (179 lesions at baseline, 173 lesions on follow-up). Table 2 summarizes intra- and interobserver variability of the longest diameter measurement of individual targets at baseline and percent change at first follow-up. CCC was high for both intra- and interobserver comparisons at baseline and follow-up [0.8602–0.9961]. The 95% limits of agreement of two (intraobserver) measurements were (−10.30%, 9.24%) at baseline and (−10.73%, 10.96%) for percent change on follow-up; interobserver were (−20.84%, 21.79%) at baseline and (−20.79, 18.71%) on follow-up.

We further investigated the impact of anatomic location of lesions on measurement variability. Table 3 summarizes intra- and interobserver measurement variability according to lesion location. Variability is visually demonstrated in Bland-Altman plots (Figure 3), with notation of anatomic location. In multivariate linear regression models, lesion location was not significant in measurement variability after controlling for intraobserver effect (p=0.35), however, it had a significant impact on measurement variability after controlling for interobserver effect (p=.008). Compared to lung as a reference, liver significantly contributed to measurement variability in addition to the interobserver variability (p=0.048) while the other locations did not.

CT Attenuation Measurements

Baseline CT attenuation (HU) measurements were performed in 54 patients with 136 targets, in patients who underwent contrast-enhanced CT at baseline. Follow-up CT attenuation measurements were performed in 44 patients with 103 targets, in patients with contrast-enhanced CT at baseline and follow-up. Three patients had lung lesions measuring less than 0 HU on follow-up, whose mean CT attenuation calculations were considered not evaluable and excluded from the analysis. Therefore, analysis of CT attenuation on follow-up was performed in 41 patients.

CT Attenuation Measurements per Patient—Table 4 summarizes intraobserver and interobserver variability of mean CT attenuation measurements (HU) per patient at baseline and percent change at follow-up CT. While CCC was high (0.9229–0.9946) for baseline measurements and percent change at follow-up, the range of 95% limits of agreement were relatively wide for percent change at follow-up, for intraobserver comparisons (−18.34%,

16.70%) as well as for interobserver (−15.31, 13.91). A 15% change in mean CT attenuation used as a response cutoff in the Choi criteria is within these 95% limits.

CT Attenuation Measurements per Lesion—Table 5 summarizes intraobserver and interobserver agreement in mean CT attenuation measurements (HU) per lesion at baseline and percent change on follow-up. CCCs were high across the comparisons (0.9033–0.9853), however, 95% limits of agreements were relatively wide for percent change at follow-up for intraobserver comparisons (−19.65%, 19.45%), as well as for baseline measurements and percent change at follow-up for interobserver comparisons ((−20.03%, 18.22%), (−24.06%, 23.16%), respectively).

DISCUSSION

High intraobserver and interobserver agreements were observed in SLD measurements and the percentage change in SLD at first follow-up in mRCC patients treated in Phase II and III trials of VEGF-targeted agents. CCCs were greater than 0.9 in nearly all assessments. A 10% tumor shrinkage threshold, which has been shown to be a predictor of survival in mRCC patients (11, 15), was outside the 95% limits of agreement of intraobserver variability assessment at baseline and for percent change in SLD at follow-up. However, 10% tumor shrinkage was within 95% limits of agreement in the interobserver assessment. To the authors' knowledge, this study represents the first comprehensive examination of intraobserver and interobserver variation in CT size and attenuation measurements in mRCC patients treated with VEGF-targeted therapies.

Prior reports have found 10% tumor shrinkage indicative of a survival benefit in mRCC patients, either using a single radiologist observer (15) or using independent central review (11). The present study is concordant with these prior reports, as we have demonstrated changes beyond (−7.30%, 7.86%) in tumor burden are true, detectable changes in overall tumor size rather than measurement error when baseline and follow-up measurements are performed by one radiologist. The findings of this study also have implications on the assessment of progressive disease, defined as at least 20% increase in tumor size from nadir (or new lesions) according to RECIST. Our findings indicate that this degree of increase represents a true increase in tumor burden as measured by one or more observers. This is important because progression-free survival is a common primary endpoint of clinical trials in this population, and our study suggests that this endpoint is a robust and reproducible one.

Our study demonstrates that 10% tumor shrinkage is within the 95% limits of agreement for interobserver variability, indicating that the 10% tumor shrinkage response threshold is within the range of measurement error using two radiologist observers. It is known that interobserver variation in tumor measurements is greater than intraobserver variation, and multiple prior reports advocate the use of a single observer or independent review committees to improve consistency in serial measurements in any one patient (24–30). Our study supports the use of a single, experienced radiologist for serial scan measurements to limit measurement variability when using a smaller change in tumor size to define response. The results are most applicable to radiologists who commonly interpret cancer imaging studies and measure target lesions in their practice.

The 10% tumor shrinkage alternative response criterion represents added value to the conventional RECIST, since RECIST clinical benefit (Partial Response + Stable Disease) has not been found to be a surrogate for treatment benefit in terms of survival (11, 15). In this setting, 10% tumor shrinkage in response to treatment is a threshold indicating true change in tumor size with prognostic value. Oncologists can use 10% tumor shrinkage early in the treatment course, to continue an effective drug with confidence. While treatment changes may not be based on whether or not a patient achieves 10% tumor shrinkage on first follow-up CT, the prognostic value of this threshold may be of use in individual cases.

It is the impression and experience of the authors that for the majority of the target lesions, particularly lymph nodes, lung, pleural, and peritoneal/retroperitoneal masses, IV contrast did not contribute significantly to the measurement variability due to marked attenuation differences between these masses and adjacent tissues (fat, air, etc.), in the presence or absence of IV contrast. In our study cohort, the majority of patients received IV contrast, (54 of 71 at baseline, 44 of 69 at follow-up), and it may not possible to determine the effect of IV contrast on inter and intra-observer variability because of lack of power. However, in a preliminary analysis to explore this point, we performed CCC analysis for baseline SLD measurements of 54 patients with contrast versus 17 patients without contrast on baseline scans, and both groups have similarly high CCCs.

The 95% limits of agreement for measurement variation were slightly wider for individual lesions than in the patient-based analyses, which is not an unexpected finding. Measuring multiple target lesions to best determine response has been advocated in prior reports (31–33). Measurement of up to 10 targets is permitted according to RECIST 1.0 while up to 5 targets are accepted according to RECIST 1.1, and almost perfect agreement in response assessment has been shown when using 1.0 versus 1.1 (34–35). In many instances in routine care, as few as three target lesions may be used in individual patients and considered reasonable in terms of response assessment and measurement reproducibility.

We performed a lesion-based analysis to evaluate the influence of anatomic location on measurement variability, which has not been previously systematically investigated. For intraobserver comparisons, 95% limits of agreement were narrowest for discrete pleural masses, followed by lung, and greatest in liver and retroperitoneum. For interobserver comparisons, 95% limits of agreement were narrowest for lung targets, followed by pleural masses, and widest in retroperitoneum and liver. Since lung lesions have relatively narrow 95% limits of agreement for both intraobserver and interobserver variation, selection of lung targets may be favorable if lung and other targets of similar size are available. Target selections are made taking into account several factors, including lesion size, reproducibility, conspicuity compared to adjacent tissues and suitability for follow-up.

In the multivariate linear regression analysis, liver location of targets contributed significantly to interobserver variability while others did not. Many liver lesions have ill-defined margins on contrast-enhanced CT and were difficult to precisely measure, likely explaining this observation.

In terms of mean CT attenuation measurements per patient and lesion, intraobserver and interobserver variations were wider. In many cases, 95% limits of agreement exceeded ±15%. These findings indicate that a 15% decrease in mean CT attenuation may be due to measurement variation rather than true change. Since Choi criteria include 15% decrease in mean CT attenuation as a composite cutoff, our results may explain why multiple reports have resulted in different conclusions on the utility of Choi criteria in assessing response in mRCC (11, 15–17). A threshold greater than 15% decrease in CT attenuation is needed to reliably indicate response.

The limitations of the present study include the relatively small number of patients scanned at a single institution. However, the number is not substantially different from other studies of alternative response criteria in this setting (11–17). Not all patients were able to receive IV contrast, limiting the size of the CT attenuation assessment cohort; this is representative of the mRCC population, including many patients with advanced age and/or a single kidney. Another limitation is that we were not able to assess the measurement variability that is inherent in conducting CT scans, in this cohort of metastatic RCC patients treated in clinical trials who did not undergo same-day repeat CT scans. This important issue has been addressed by a recent study in which CT scans were repeated within 15 minutes and target lesions were measured by three radiologists (21, 31). We also did not assess advanced image processing techniques, such as volumetric measurements, since these measurements are not routinely used in clinical practice in determining response assessment of mRCC patients, and because our central objective was to demonstrate that 10% tumor shrinkage, a known predictor of survival, is a reliable marker in terms of measurement reproducibility. We designed the study to utilize size and/or CT attenuation measurements obtained with clinical CT scans which patients typically undergo in routine oncologic care.

In conclusion, radiologic response, defined by 10% decrease in SLD of target lesions in mRCC patients treated with VEGF-targeted therapies, is reliable and reproducible when evaluated by a single radiologist observer. The anatomic location of the lesions has a significant impact on measurement variability, with liver location contributing to the measurement variability greatest after controlling for interobserver variation. Use of a single, trained observer for both baseline and follow-up measurements is recommended when assigning radiologic response using the alternative, “10% tumor shrinkage threshold” to better predict clinical outcome.

Acknowledgments

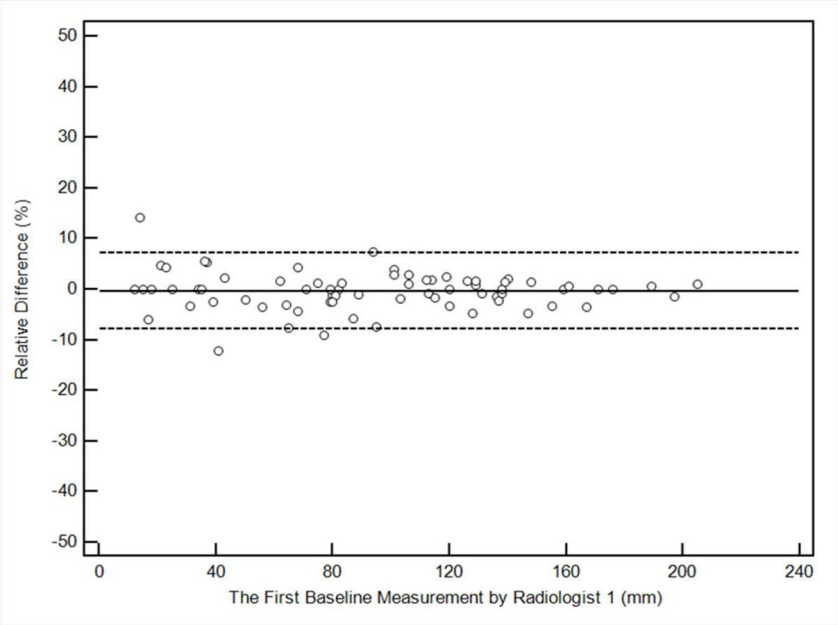
Acknowledgment of Research Support: Association of University Radiologists-GE-Radiology Research Academic Fellowship-K.M.K; 1K23CA157631 (NCI/M.N.)

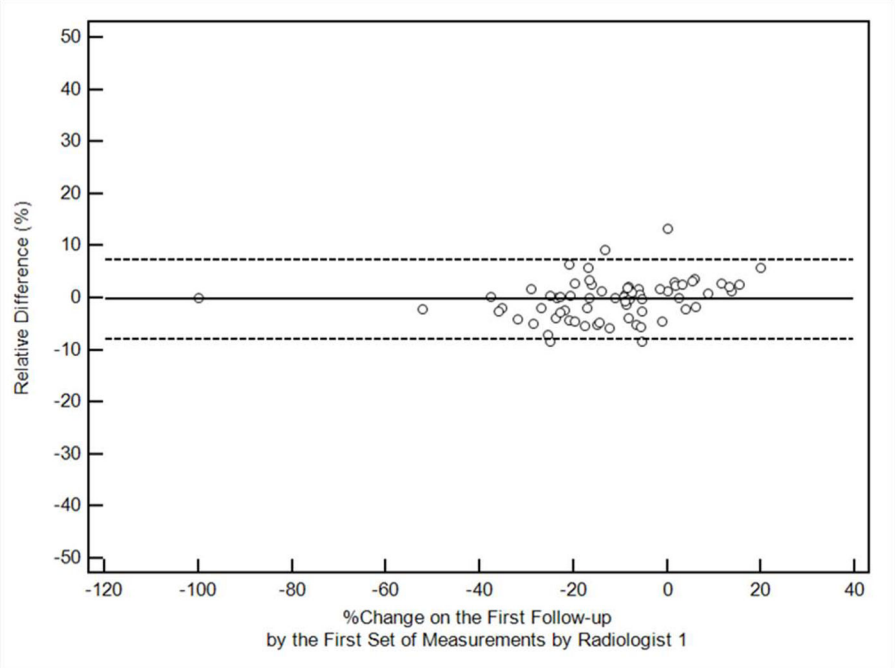
References

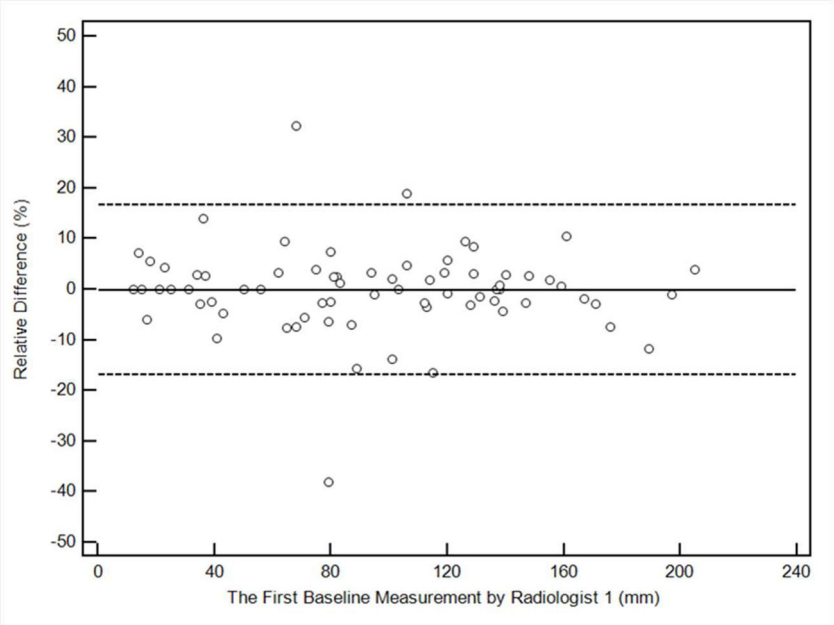
1. American Cancer Society. Cancer facts and figures. 2012. http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-031941.pdf
2. Rabinovich RA, Zelefsky MJ, Gaynor JJ, Fuks Z. Patterns of Failure Following Surgical Resection of Renal Cell Carcinoma: Implications for Adjuvant Local and Systemic Therapy. J Clin Onc. 1994; 12(1):206–212.

3. Janzen NK, Kim HL, Figlin RA, Belldegrun HA. Surveillance after radical or partial nephrectomy for localized renal cell carcinoma and management of recurrent disease. Urol Clin N Am. 2003; 30:843–852.
4. Motzer RJ, Bukowki RM. Targeted Therapy for Metastatic Renal Cell Carcinoma. J Clin Oncol. 24:5601–5608. [PubMed: 17158546]
5. Yang JC, Haworth L, Sherry RM, et al. A Randomized Trial of Bevacizumab, an Anti-Vascular Endothelial Growth Factor Antibody, for Metastatic Renal Cell Carcinoma. N Engl J Med. 2003; 349:427–434. [PubMed: 12890841]
6. Motzer RJ, Hutson TE, Tomczak P, et al. Sunitinib versus Interferon Alfa in Metastatic Renal-Cell Carcinoma. N Engl J Med. 2007; 356:115–124. [PubMed: 17215529]
7. Escudier B, Eisen T, Stadler WM, et al. TARGET Study Group. Sorafenib in Advanced Renal Cell Carcinoma. N Engl J Med. 2007; 356:125–134. [PubMed: 17215530]
8. Motzer RJ, Michaelson MD, Rosenberg J, et al. Sunitinib Efficacy Against Advanced Renal Cell Carcinoma. J Urol. 2007; 178:1883–1887. [PubMed: 17868732]
9. 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): 205–16. [PubMed: 10655437]
10. Motzer RJ, Hutson TE, Cella D, Reeves J, Hawkins R, Guo J, Nathan P, Staehler M, de Souza P, Merchan JR, Boleti E, Fife K, Jin J, Jones R, Uemura H, De Giorgi U, Harmenberg U, Wang J, Sternberg CN, Deen K, McCann L, Hackshaw MD, Crescenzo R, Pandite LN, Choueiri TK. Pazopanib versus sunitinib in metastatic renal-cell carcinoma. N Engl J Med. 2013 Aug 22; 369(8):722–31. [PubMed: 23964934]
11. Thiam R, Fournier LS, Trinquart L, et al. Optimizing the size variation threshold for the CT evaluation of response in metastatic renal cell carcinoma treated with sunitinib. Ann Oncol. 2010; 21(5):936–941. [PubMed: 19889607]
12. van der Veldt AAM, Meijerink MR, van den Eertwegh AJM, et al. Choi response criteria for early prediction of clinical outcome in patients with metastatic renal cell carcinoma treated with Sunitinib. Br J Cancer. 2010; 102:803–809. [PubMed: 20145618]
13. Smith AD, Leiber ML, Shah SN. Assessing Tumor Response and Detecting Recurrence in Metastatic Renal Cell Carcinoma on Targeted Therapy: Importance of Size and Attenuation on Contrast-Enhanced CT. AJR. 2010; 194:157–165. [PubMed: 20028918]
14. Smith AD, Shah SN, Rini BI, et al. Morphology, Attenuation, Size and Structure (MASS) Criteria; Assessing Response and Predicting Clinical Outcome in Metastatic Renal Cell Carcinoma on Antiangiogenic Targeted Therapy. AJR. 2010; 194:1470–1478. [PubMed: 20489085]
15. Krajewski KM, Guo M, Van den Abbeele AD, et al. Comparison of Four Early Posttherapy Imaging Changes (EPTIC; RECIST 1. 0, Tumor Shrinkage, Computed Tomography Tumor Density, Choi Criteria) in Assessing Outcome to Vascular Endothelial Growth Factor-Targeted Therapy in Patients with Advanced Renal Cell Carcinoma. Eur Urol. 2011; 59:856–862. [PubMed: 21306819]
16. Hittinger M, Staehler M, Schramm N, et al. Course and density of metastatic renal cell carcinoma lesions in the early follow-up of molecular targeted therapy. Urol Oncol. 2012; 30:695–703. [PubMed: 21865061]
17. Schmidt N, Hess V, Zumbrunn T, et al. Choi response for prediction of survival in patients with metastatic renal cell carcinoma treated with anti-angiogenic therapies. Eur Radiol. 2013; 23(3): 632–9. [PubMed: 22918564]
18. 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; 25:1753–1759. [PubMed: 17470865]
19. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45:255–268. [PubMed: 2720055]
20. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1:307–310. [PubMed: 2868172]

21. Zhao B, James LP, Moskowitz CS, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology. 2009; 252:263–272. [PubMed: 19561260]
22. Nishino M, Guo M, Jackman DM, et al. CT Tumor Volume Measurement in Advanced Non-small-cell Lung Cancer: Performance Characteristics of an Emerging Clinical Tool. Acad Radiol. 2011; 18:54–62. [PubMed: 21036632]
23. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20:37–46.
24. Erasmus JJ, Gladish GW, Broemeling L, 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; 21:2574–2582. [PubMed: 12829678]
25. Belton AL, Saini S, Liebermann K, et al. Tumor Size Measurement in an Oncology Clinical Trial: Comparison Between Off-site and On-site Measurements. Clin Radiol. 2003; 58:311–314. [PubMed: 12662953]
26. Tang PA, Pond GR, Chen EX. Influence of an Independent Review Committee on Assessment of Response Rate and Progression-Free Survival in Phase III Clinical Trials. Ann Oncol. 2010; 21:19–25. [PubMed: 19875758]
27. Suzuki C, Torkzad MR, Jacobsson H, et al. Interobserver and Intraobserver Variability in the Response Evaluation of Cancer Therapy According to RECIST and WHO-Criteria. Acta Oncologica. 2010; 49:509–514. [PubMed: 20397778]
28. Skougaard K, Dusgaard Mccullagh MJ, Nielsen D, et al. Observer Variability in a Phase II Trial-Assessing Consistency in RECIST Application. Acta Oncologica. 2012; 51:774–780. [PubMed: 22432439]
29. Muenzel D, Engels HP, Bruegel M, et al. Intra- and Inter-observer Variability in Measurement of Target Lesions: Implication on Response Evaluation According to RECIST 1. 1. Radiol Oncol. 2012; 46:8–18. [PubMed: 22933974]
30. Urban T, Harris GJ, Barish MA, et al. Benefits of utilizing image analysts for radiological measurements in oncology clinical trials. Applied Clinical Trials. 2010; 19(9):32–36.
31. Oxnard GR, Zhao B, Sima CS, et al. Variability of Lung Tumor Measurementson Repeat Computed Tomography Scans Taken Within 15 Minutes. J Clin Oncol. 2011; 29:3114–3119. [PubMed: 21730273]
32. Hopper KD, Kasales CJ, Van Slyke MA, et al. Analysis of Interobserver and Intraobserver Variability in CT Tumor Measurements. AJR. 1996; 167:851–854. [PubMed: 8819370]
33. Schwartz LH, Mazumdar M, Brown W, et al. Variability in Response Assessment in Solid Tumors: Effect of Number of Lesions Chosen for Measurement. Clin Cancer Res. 2003; 9:4318–4323. [PubMed: 14555501]
34. Nishino M, Jackman DM, Hatabu H, et al. New Response Evaluation Criteria in Solid Tumors (RECIST) guidelines for advanced non-small cell lung cancer: comparison with original RECIST and impact on assessment of tumor response to targeted therapy. AJR. 2010; 195:W221–W228. [PubMed: 20729419]
35. 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):228–47. [PubMed: 19097774]







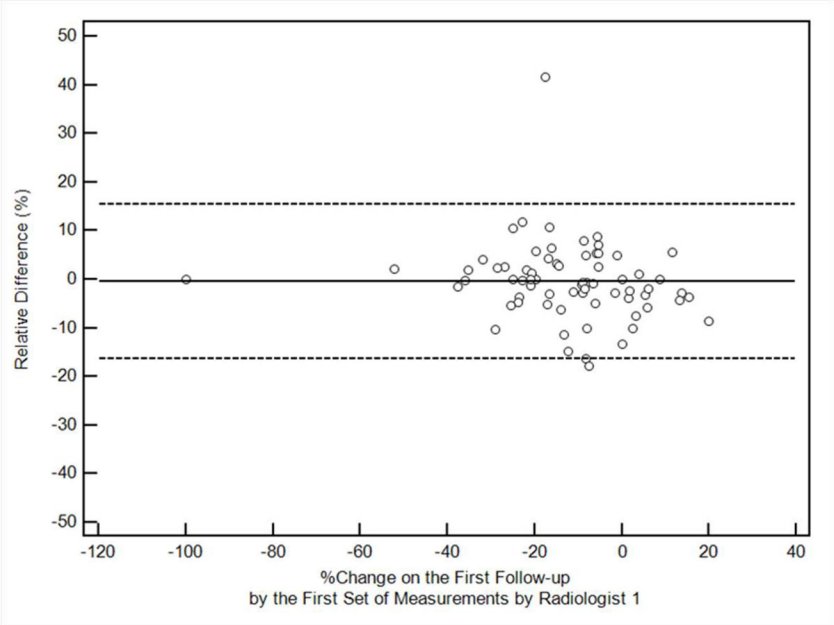


Figure 1. Bland-Altman plots for sum long axis diameter measurements on CT by two radiologists

The figures demonstrate the intraobserver or interobserver variability, by plotting the relative difference between measurements against the first measurement of Radiologist 1. The solid line represents the mean relative difference (%), and the dashed lines represent the upper and lower 95% limits of agreement.

a. Intraobserver variability at baseline assessment, comparing first and second measurements by Radiologist 1

b. Intraobserver variability of the percent changes at the first follow-up, comparing first and second measurements by Radiologist 1

c. Interobserver variability of baseline measurements, comparing the first measurement by Radiologist 1 and Radiologist 2

d. Interobserver variability at the first follow-up, comparing the first measurement by Radiologist 1 and Radiologist 2

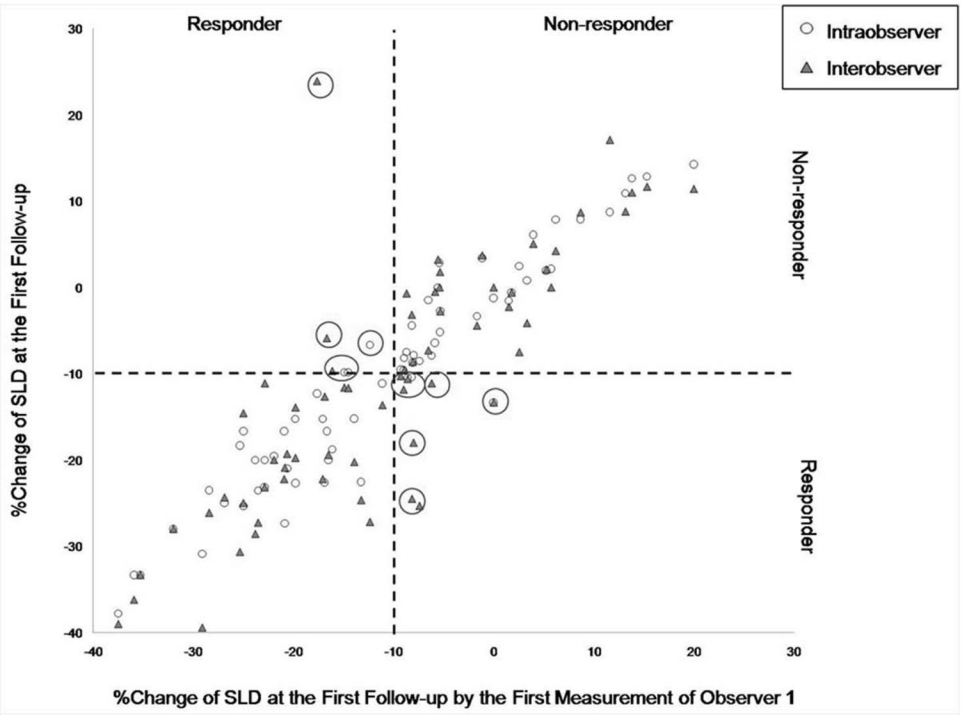


Figure 2. Scatterplot of the percent changes in SLD and response assessment using the −10% cutoff value at the first follow-up

The second set of measurements by Radiologist 1 (circles, intraobserver) and set of measurements by Radiologist 2 (triangles, interobserver) are plotted against the percent change according to the first measurement of Radiologist 1. Dashed lines represent 10% tumor shrinkage, defining responders and non-responders. Observations in the upper left and lower right (large circles) are discordant observations. (Concordant observations obtained from two patients [−52.2%, −50.0%, −50%; 1st and 2nd measurements by Radiologist 1, measurement by Radiologist 2, respectively] and [−100% for all three measurements] are not included in the figure, since the range of axis was optimized to demonstrate observations close to −10%.).

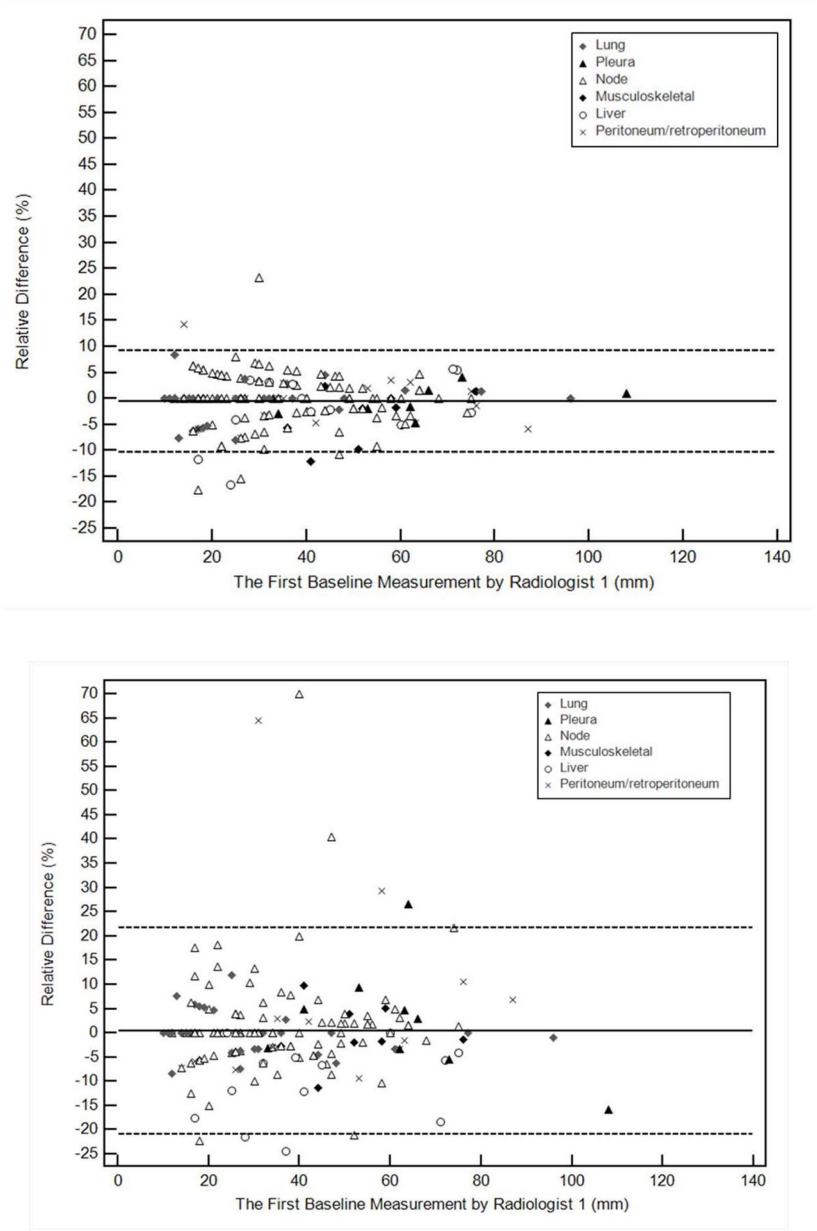


Figure 3.

Bland-Altman plots for long axis diameter measurements of individual targets at baseline by two radiologists. The figures demonstrate intraobserver or interobserver variability as a function of average measurements. Relative difference between baseline measurements is plotted by the average measurement, for the two measurements by Radiologist 1 (a, intraobserver), and for the measurements of Radiologist 1 and Radiologist 2 (b, interobserver). The solid line represents the mean relative difference (%), and the dashed lines represent the upper and lower 95% limits of agreement.

Table 1

Patient-based analysis: intraobserver and interobserver variability in the sum longest diameter (SLD) of target lesions for baseline measurements and for the percent change at the first follow-up scan.

Intraobserver variability of SLD (Two measurements by Radiologist 1)			
	CCC (95%CI)	**Mean relative difference(%)**	**95% limits of agreement(%)**
Baseline (n=71)	0.9984 (0.9974–0.9990)	−0.365	−7.90, 7.17
%change on follow-up (n=69)	0.9747 (0.9597–0.9841)	0.282	−7.30, 7.86
Interobserver variability of SLD (Measurements by Radiologist 1 vs. Radiologist 2)			
	CCC (95%CI)	Mean relative difference(%)	95% limits of agreement(%)
Baseline (n=71)	0.9875 (0.9800–0.9921)	−0.173	−17.0, 16.6
%change on follow-up (n=69)	0.8969 (0.8389–0.9348)	−0.444	−16.3, 15.4

Table 2

Lesion-based analysis: intraobserver and interobserver variability in the longest diameter measurement of individual target lesions for the baseline measurements and for the percent change at first follow-up CT.

Intraobserver variability of measurements per lesion (Two measurements by Radiologist 1)			
	CCC (95%CI)	**Mean relative difference(%)**	**95% limits of agreement(%)**
Baseline (n=179 lesions)	0.9961 (0.9947–0.9971)	−0.528	−10.30, 9.24
%change on follow-up (n=173 lesions)	0.9536 (0.9381–0.9653)	0.118	−10.73, 10.96
Interobserver variability of measurements per lesion (Measurements by Radiologist 1 vs. Radiologist 2)			
	CCC (95%CI)	Mean relative difference(%)	95% limits of agreement(%)
Baseline (n=179 lesions)	0.9712 (0.9616–0.9784)	0.474	−20.84, 21.79
%change on follow-up (n=173 lesions)	0.8602 (0.8162–0.8942)	−1.036	−20.79, 18.71

Table 3

Lesion-based sub-analysis: intraobserver and interobserver variability in longest diameter measurement at baseline, categorized by lesion location.

	Intraobserver Variability			Interobserver Variability		
Location of the lesions	**CCC (95%CI)**	**Mean relative difference (%)**	**95% limits of agreement(%)**	**CCC (95%CI)**	**Mean relative difference (%)**	**95% limits of agreement(%)**
1: Lung (n=33)	0.9991 (0.9983–0.9996)	−0.347	−7.25, 6.55	0.9982 (0.9964–0.9991)	−0.323	−8.92, 8.27
2: Pleura (n=11)	0.9976 (0.9918–0.9993)	−0.503	−5.43, 4.42	0.9337(0.7858–0.9806)	1.93	−18.7, 22.5
3: Node (n=97)	0.9941 (0.9913–0.9960)	−0.201	−10.23, −10.63	0.9621 (0.9447–0.9740)	1.45	−19.7, 22.6
4: MSK (n=8)	0.9718 (0.8873–0.9932)	−3.45	−13.77, 6.87	0.9744 (0.8805–0.9947)	−0.042	−12.48, 12.39
5: Liver (n=13)	0.9931 (0.9764–0.9980)	−1.86	−14.79, 11.07	0.9601 (0.8911–0.9857)	−10.27	−26.03, 5.94
6: Peritoneum/retro (N=17)	0.9963 (0.9903–0.9986)	−0.37	−10.44, 9.69	0.9514 (0.8779–0.9818)	3.98	−31.41, 39.38

Table 4

Patient-based analysis using the average CT attenuation of up to 3 lesions/patient: intraobserver and interobserver variability in mean CT attenuation measurements (HU)/patient at baseline and percent change in mean attenuation (HU) at follow-up.

Intraobserver variability of CT attenuation measurements			
	CCC (95%CI)	**Mean relative difference(%)**	**95% limits of agreement(%)**
Baseline (n=54 patients)	0.9946 (0.9907–0.9968)	−0.29	−6.74–6.17
%change on follow-up (n=41)	0.9229 (0.8604–0.9580)	−0.82	−18.34, 16.70
Interobserver variability of CT attenuation measurements			
	CCC (95%CI)	Mean relative difference(%)	95% limits of agreement(%)
Baseline (n=54)	0.9752 (0.9578–0.9855)	0.59	−12.92–14.1
%change on follow-up (n=41)	0.9499 (0.9101–0.9723)	−0.70	−15.31, 13.91

Table 5

Lesion-based analysis: intraobserver and interobserver variability in mean CT attenuation measurements (HU) at baseline, according to anatomic location.

Intraobserver variability of CT attenuation measurements			
	CCC (95%CI)	**Mean relative difference(%)**	**95% limits of agreement(%)**
Baseline CT attenuation (n=136 lesions)	0.9853 (0.9795–0.9895)	0.010	−10.85, 10.87
%change on 1st follow-up scan (n=103 lesions)	0.9300 (0.8983–0.9521)	−0.100	−19.65, 19.45
Interobserver variability of CT attenuation measurements			
	CCC (95%CI)	Mean relative difference(%)	95% limits of agreement(%)
Baseline CT attenuation (n=136 lesions)	0.9496 (0.9300–0.9638)	−0.921	−20.03, 18.22
%change on 1st follow-up scan (n=103 lesions)	0.9033 (0.8609–0.9332)	−0.475	−24.06, 23.16