



Variability of quantitative measurements of metastatic liver lesions: a multi-radiation-dose-level and multi-reader comparison

Yuqin Ding^{1,2} · Daniele Marin¹ · Federica Vernuccio³ · Fernando Gonzalez⁴ · Hannah V. Williamson⁵ · Hans-Christoph Becker⁶ · Bhavik N. Patel⁶ · Justin Solomon⁷ · Juan Carlos Ramirez-Giraldo⁸ · Ehsan Samei⁷ · Rendon C. Nelson¹ · Mathias Meyer¹

Received: 15 April 2020 / Accepted: 26 May 2020 / Published online: 10 June 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Purpose To evaluate the variability of quantitative measurements of metastatic liver lesions by using a multi-radiation-dose-level and multi-reader comparison.

Methods Twenty-three study subjects (mean age, 60 years) with 39 liver lesions who underwent a single-energy dual-source contrast-enhanced staging CT between June 2015 and December 2015 were included. CT data were reconstructed with seven different radiation dose levels (ranging from 25 to 100%) on the basis of a single CT acquisition. Four radiologists independently performed manual tumor measurements and two radiologists performed semi-automated tumor measurements. Interobserver, intraobserver, and interdose sources of variability for longest diameter and volumetric measurements were estimated and compared using Wilcoxon rank-sum tests and intraclass correlation coefficients.

Results Inter- and intraobserver variabilities for manual measurements of the longest diameter were higher compared to semi-automated measurements ($p < 0.001$ for overall). Inter- and intraobserver variabilities of volume measurements were higher compared to the longest diameter measurement ($p < 0.001$ for overall). Quantitative measurements were statistically different at $< 50\%$ radiation dose levels for semi-automated measurements of the longest diameter, and at 25% radiation dose level for volumetric measurements. The variability related to radiation dose was not significantly different from the inter- and intraobserver variability for the measurements of the longest diameter.

Conclusion The variability related to radiation dose is comparable to the inter- and intraobserver variability for measurements of the longest diameter. Caution should be warranted in reducing radiation dose level below 50% of a conventional CT protocol due to the potentially detrimental impact on the assessment of lesion response in the liver.

Keywords Dual-source computed tomography · Radiation dose · Intraobserver variability · Interobserver variability · Liver metastasis

Abbreviations

CI Confidence interval
ICC Intraclass correlation coefficient

SAFIRE Sinogram-affirmed iterative reconstruction
CTDI_{vol} Volume CT dose index

✉ Mathias Meyer
mathias.meyer@duke.edu

¹ Department of Radiology, Duke University Medical Center, 2301 Erwin Rd, Durham, NC 27710, USA

² Department of Radiology, Zhongshan Hospital, Fudan University; Shanghai Institute of Medical Imaging, Shanghai 200032, China

³ Department ProMISE (Department of Health Promotion, Mother and Child Care, Internal Medicine and Medical Specialties), University Hospital of Palermo, Piazza Delle Cliniche, 90127 Palermo, Italy

⁴ Department of Radiology, Clínica Alemana de Santiago, Universidad del Desarrollo, 8320000 Santiago, Chile

⁵ Duke Cancer Institute-Biostatistics, Duke University Medical Center, Durham, NC 27705, USA

⁶ Department of Radiology, Stanford University, School of Medicine, Stanford, CA 94305, USA

⁷ The Carl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University Medical Center, Durham, NC 27705, USA

⁸ Siemens Healthineers, 40 Liberty Boulevard, Malvern, PA 19355, USA

Introduction

In oncology imaging, size-based assessment of tumor burden remains the standard objective metric of tumor response to therapy for both clinical trials and daily clinical practice [1–3]. Ever since the introduction of the RECIST1.0 and RECIST1.1 guidelines [1, 2], multiple studies have investigated inter- and intraobserver agreement of RECIST measurements. The majority of these studies focused on the effect of manual versus semi-automated diameter and volumetric measurements of target lesions [4–13]. Radiologist-dependent factors, such as variations in target lesions selection, annotation precision, and working experience level, represent a major source of variability in the assessment of tumor response [13, 14]. However, the impact of CT acquisition and reconstruction as sources of variability in tumor response assessment are often undervalued. Previous studies have shown non-negligible variability in quantitative measurements of lung nodules with changes in section thickness, section interval as well as reconstruction algorithm [15–17]. In fact, the complex interplay of different CT systems and variances in scanning parameters at either the same or different institution has led to substantial variability in radiation dose levels with up to 42% variability in the same patient undergoing identical repeat thoracoabdominal staging CT protocols [18, 19].

A previous study has demonstrated that the volumetric quantification of lung nodule is extremely robust even at radiation dose levels as low as 25%, 10%, and 3% of a clinical standard protocols [20]. While radiation dose may not impact volumetric lung nodule quantification, likely due to the high lesion-to-background contrast ratio between solid pulmonary nodules and the lung parenchyma, it remains to be determined whether a similar conclusion can be applied to lesions of soft tissue organs (e.g., liver) with low inherent lesion-to-background contrast. We postulate that radiation dose could be a source of measurement variability for focal liver lesions. Thus, the purpose of this study was to assess the variability related to radiation dose for manual and semi-automated measurements of metastatic liver lesions, by using a multi-radiation-dose-level and multi-reader comparison.

Materials and methods

This was a secondary analysis of data from a Health Insurance Portability and Accountability Act–compliant, institutional review board-approved prospective trial designed to assess the effect of iterative reconstruction and radiation

dose reduction in abdominal CT. This study was approved by the institutional review board of Duke University and a waiver of written informed consent was obtained. One author of the study (J.C.R.G.) is an employee of Siemens Healthineers, two authors (D.M. and M.M.) received research support (provision of software tools used in this study) from Siemens Healthineers and MintLesion. All other authors are not employees of or consultants for the industry and had control of any data or information that might present a conflict of interest.

Study participants

Study participants were eligible for inclusion between June 2015 and December 2015 if: (i) they underwent a single-energy dual-source contrast-enhanced staging CT during the portal venous phase and (ii) were known or suspected to have liver metastases from colon cancer on the basis of the results of: (iia) previous multidetector CT or ultrasonographic examination findings or (iib) increased carcinoembryonic antigen tumor marker levels (> 5 ng/mL). Study participants were excluded if: (i) the study participants total body weight was greater than 118 kg (260 lbs), which is our departmental cut-off for performing a dual-energy CT examination, (ii) there were metal artifacts from the spine or abdominal clips/stents affected the liver parenchyma, and/or (iii) image quality was non-diagnostic due to increased noise texture and contrast media timing. 23 patients (12 men and 11 women; mean age, 60 years, age range 28–73 years) fulfilled our inclusion criteria and represent the final subject cohort.

CT parameters and image reconstruction

All scans were performed by using a second-generation dual-source multidetector CT platform (Somatom Definition Flash; Siemens Healthineers, Forchheim, Germany). Each X-ray tube (tube A and tube B) had a tube voltage of 120 kVp but different tube currents (in milliamperes). By distributing the tube current unevenly between Tube A and Tube B (at the same tube voltage), it is possible to reconstruct images as if they had been acquired at any radiation dose in the range of $\min(D_A, D_B)$ to $D_A + D_B$ [21]. This was done by linearly combining the two image sets according to a weighting factor as follows:

$I_{A+B,w} = wI_A + (1 - w)I_B$, where $I_{A+B,w}$ is the combined image, I_A and I_B are images from Tube A and Tube B, respectively, and w is a weighting factor with values between 0 and 1. Choosing different values of w allows one to produce images corresponding to different dose levels according to the following relationship.

$D_{A+B,w} = \frac{D_A}{w^2 + (1-2w+w^2)\frac{D_A}{D_B}}$, where $D_{A+B,w}$ represents the noise equivalent dose for the combined image and D_A and D_B are the dose from Tube A and Tube B, respectively.

All patients were positioned supine on the scanning table and underwent scanning in the craniocaudal direction. Each patient underwent a routine-dose single-breath-hold acquisition of the chest, abdomen, and pelvis during the portal venous phase (approximately 70 s [range 62–83 s] after the start of contrast medium injection) by using a single-energy dual-source CT acquisition. Detailed CT scanning parameters were as follows: tube potential, 120 kVp; tube current modulation, CareDose4D (Siemens Healthineers, Forchheim, Germany); quality reference effective milliamperere second (tube A = 150 mAs, tube B = 50 mAs); detector configuration, $2 \times 64 \times 0.6$ mm; rotation time, 0.33 s; helical pitch, 0.8; section thickness, 5 mm. All patients received 150 mL of an intravenous nonionic contrast medium (iopamidol, Isovue 300; Bracco Diagnostics, Princeton, NJ) through an 18–20 gauge angiocatheter inserted into an antecubital fossa vein by using a dual-chamber mechanical power injector (Empower; E-Z-Em, Bracco Diagnostics, Princeton, NJ) at a flow rate of 3 mL/s. To obtain images during the hepatic venous phase, a bolus tracking technique was used. A monitoring region-of-interest (ROI) was placed in the liver parenchymal and scan acquisition was triggered once an enhancement threshold of +45–50 Hounsfield units (HU) was reached. There was an additional 3 s delay from triggering to scan acquisition.

The raw projection data from each of the CT examinations were anonymized and exported from the scanner console into an external hard drive and transferred to an offline prototype research reconstruction engine (ReconCT, Siemens Healthineers, Forchheim, Germany). All images were reconstructed using an image matrix of 512×512 pixels. Seven different radiation dose levels (100%, 87.5%, 75%, 62.5%, 50%, 37.5%, 25%) were reconstructed using the above-mentioned equations and a dedicated second-generation iterative reconstruction algorithm at strength 2 out of 5 (SAFIRE, Siemens Healthineers, Forchheim, Germany) using a dedicated soft tissue kernel (I30) and a section thickness of 5.0 mm (increment 4.0 mm). For each patient, the volume CT dose index (CTDIvol), dose length product (DLP), effective diameter, and size-specific dose estimate (SSDE) were recorded.

Image interpretation

The reconstructed radiation dose level data sets of each patient were randomly divided into seven groups and evaluated by four radiologists independently (D.M., F.G., M.M., and F.V., having 10, 7, 6, and 4 years of experience in liver imaging, respectively). Target lesions were selected in

accordance with RECIST 1.1 by an experienced radiologist (D.M., with 10-year experience in RECIST 1.1 CT response assessment). Each target lesion was individually presented to each radiologist in a separate read on a dedicated offline image viewer (Horos v3.3.5) and the radiologist performed a measurement according to RECIST 1.1 [2]. For each target lesion presentation, the slice position, window width and levels were kept constant across all radiologists using a preset soft tissue window (window width, 350 HU; window level, +40 HU).

Semi-automated measurements were taken independently by two radiologists (F.G. and M.M.) using two dedicated semi-automated oncology software tools, MintLesion (version 2.6.4; MintMedical, Heidelberg, Germany) and MM Oncology (Syngo.via, Version VB30, Siemens Healthineers, Forchheim, Germany). An example of semi-automated measurement of the longest diameter and volume of liver lesion across all dose levels is shown in Fig. 1. The observers started the semi-automated segmentation process by drawing a circle/diameter over the target lesions on the axial image using the same slice position as the manual measurements were performed. The longest diameter and volume of the liver lesion were automatically calculated by the two different software tools. Each segmented target lesion was verified visually by each observer and manual adjustments could be made with dedicated correction tools if the lesion was not fully included in the segmentation field or if there was too much overlap of adjacent tissue.

Observers were asked to specify an observer confidence score by using a scale from 0 to 100 according to the recommendations of the European Guidelines on Quality Criteria for Computed Tomography [22], where a score of 100 indicated the greatest observer's confidence level for measurement of liver lesions. Intraobserver variability was assessed by each observer during a second reading session which was 30 days apart from the first reading session. To minimize the effect of recall bias from the interpretation of multiple data sets within the same patient [23], image data sets were presented in a different randomized order than the first read for each observer.

Statistical analysis

The absolute and percentage differences were calculated from predefined references' quantitative measurements. The senior observer served as a reference for the manual quantitative measurement and MintLesion software for the semi-automated quantitative measurement to evaluate interobserver variability. The first quantitative measurement of each lesion for each observer served as a reference to evaluate intraobserver variability. Similarly, the quantitative measurement at 100% dose level served as a reference to evaluate dose variability.

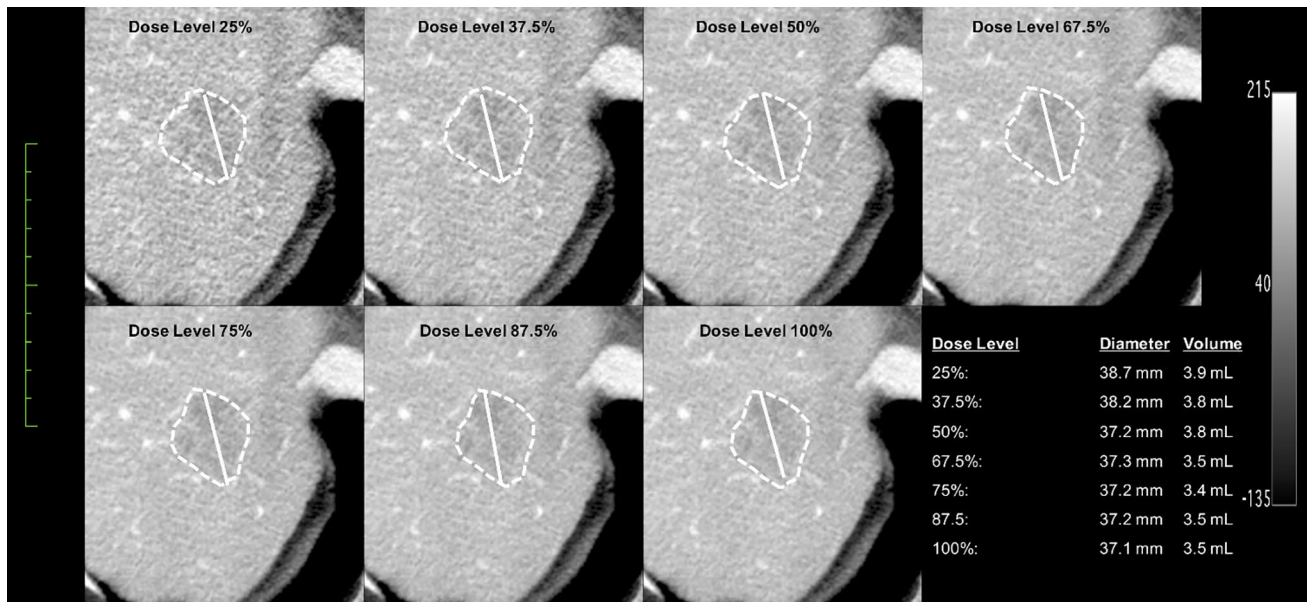


Fig. 1 An example of semi-automated measurement of the longest diameter and volume of liver lesion across all dose levels. Absolute percentage change of measurements at 25% dose level compared to

100% dose level is 4.3% for the longest diameter and 11.4% for the volume of liver metastasis

Wilcoxon rank-sum tests were used to compare the interobserver, intraobserver, and radiation dose variability, using the metrics of absolute difference and absolute percentage difference from the respective reference values. Due to the large number of comparisons, a Bonferroni correction was implemented for the pairwise comparisons of interobserver-, intraobserver-, and dose variability, with $p < 0.0008$ considered statistically significant. One-way random-effects intraclass correlation coefficients (ICC) and associated 95% confidence intervals (CI) were also used to quantify the level of agreement across all measurements of a particular type within each dose level. ICCs were calculated using the “irr” package in R [24]. We considered ICC < 0.50 poor, 0.50 – 0.75 moderate, 0.75 – 0.90 good, and > 0.9 , excellent [25].

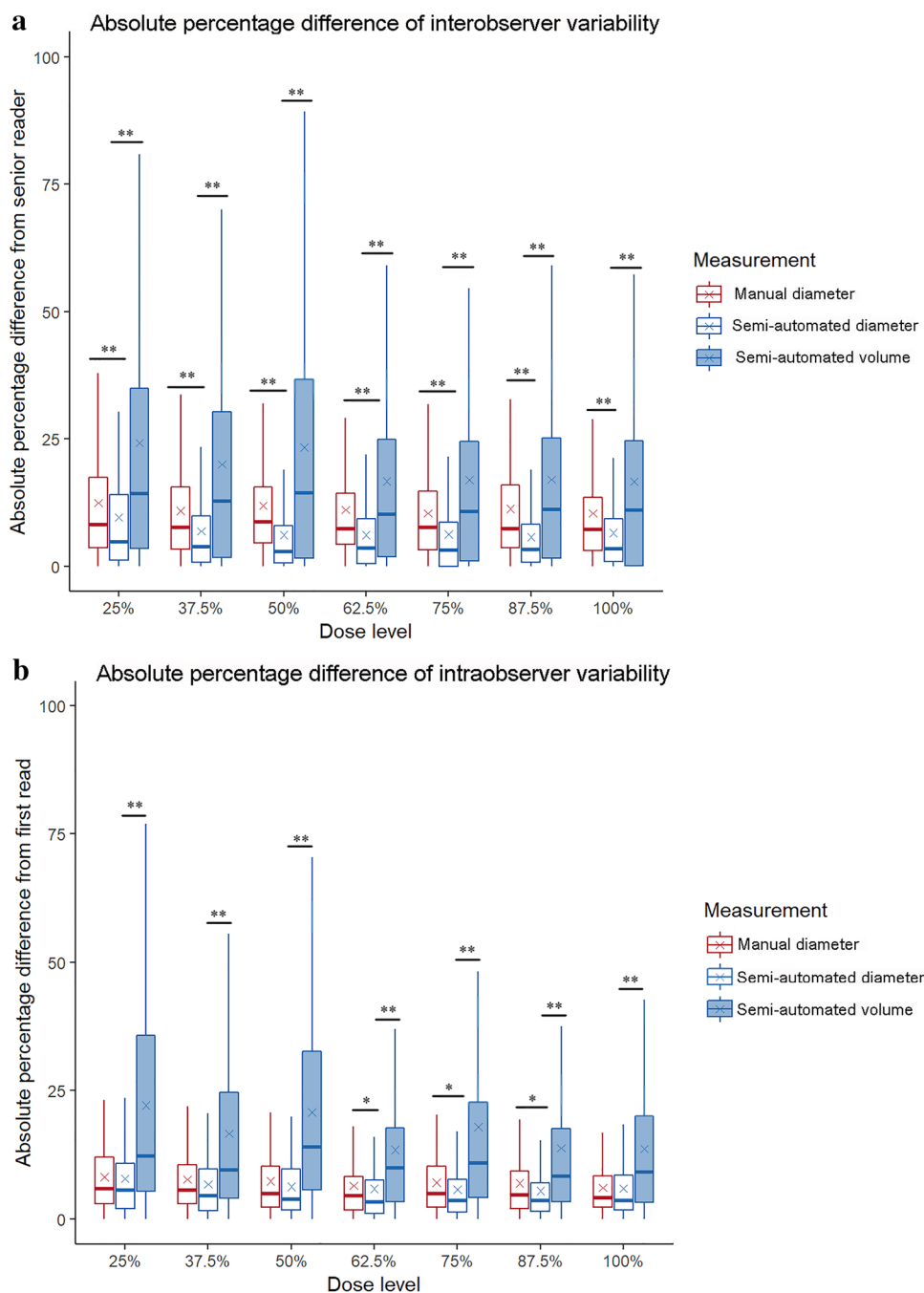
The impact of radiation dose level on quantitative measurement and observer confidence was analyzed by fitting a linear mixed-effects regression model. In the quantitative measurement model, dose was a fixed effect, with random intercepts for individual readers and lesions. In the confidence model, observer (manual versus semi-automated), dose, and the interaction between observer and dose were treated as fixed effects, with random intercepts for individual readers and lesions. All statistical analyses were performed with R software, version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria).

Results

A total of 23 patients (mean age 60 ± 12 years, range 28–73 years and mean effective diameter 28.7 ± 3.4 cm, range 23.7–34.5 cm) including 12 men (mean age 60 ± 10 years, range 35–72 years and mean effective diameter 29.1 ± 3.8 cm, range 23.8–34.5 cm) and 11 women (mean age 61 ± 13 years, range 28–73 years and mean effective diameter 28.3 ± 3.1 cm, range 23.7–34.5 cm), with 39 metastatic liver lesions were included in our study. The manual and semi-automated measurements of the longest diameter and volume at 100% dose level were on average 2.7 ± 1.4 cm (range 1.0–8.1 cm), 2.8 ± 1.6 cm (range 1.0–9.8 cm), and 13.9 ± 25.8 cm³ (range 0.52–142.13 cm³), respectively.

The median absolute percentage differences between observers (interobserver variabilities) were 7.7% (IQR 3.7–15.3%) for manual longest diameter, 3.6% (IQR 0.7–9.3%) for semi-automated longest diameter, and 12.5% (IQR 1.8–30.4%) for volumetric measurements. The median absolute percent differences between reads (intraobserver variabilities) were 5.0% (IQR 2.3–9.7%) for manual longest diameter, 3.9% (IQR 1.5–8.6%) for semi-automated longest diameter, and 10.7% (IQR 4.2–24.3%) for volumetric measurements. The inter- and intraobserver variability of the manual measurements of the longest

Fig. 2 Box plots show interobserver variability (a), intraobserver variability (b), and dose variability (c). The variability related to radiation dose was not significantly different from inter- and intraobserver variability for both manual and semi-automated measurements of the longest diameter. The variability related to radiation dose for volume measurements was statistically higher than that of inter- and intraobserver variability at 50% and 100% dose levels. * p value < 0.05, ** p value < 0.001. p values are from Wilcoxon rank-sum tests



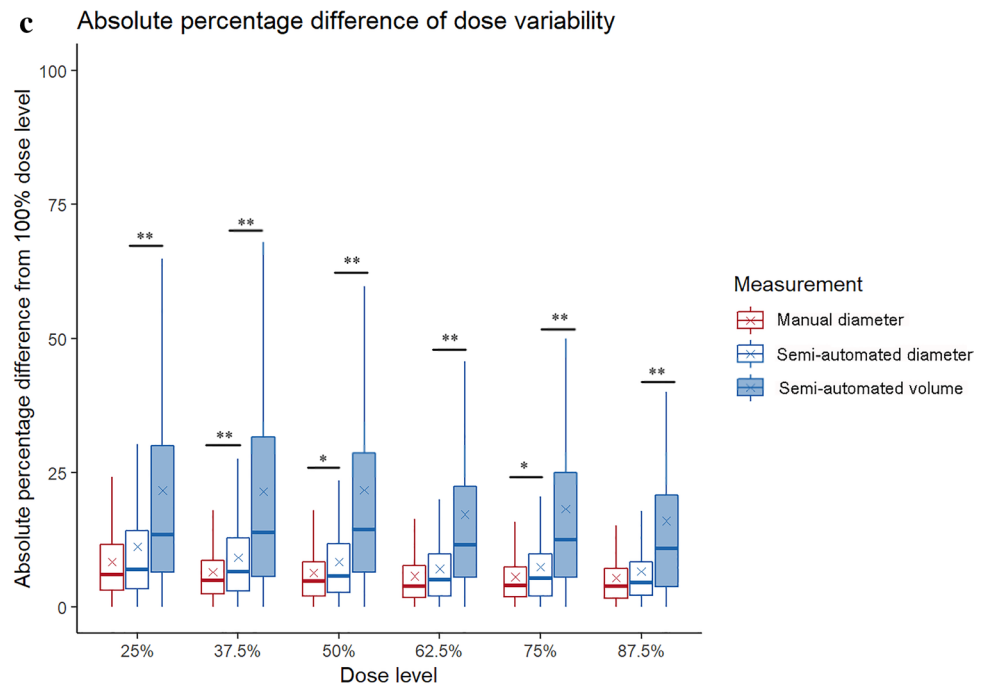
diameter were higher (though not all statistically significant at all dose levels) than that of the semi-automated measurements ($p < 0.001$ for overall). The inter- and intraobserver variability were higher for the volumetric measurement, compared to both manual and semi-automated measurements of the longest diameter ($p < 0.001$ for overall; Fig. 2a, b).

The median absolute percentage differences between dose levels (dose variabilities) were 4.5% (IQR 1.9–8.6%) for manual longest diameter, 5.6% (IQR 2.3–11.4%) for semi-automated longest diameter and 13.6% (IQR 5.7–28.4%) for

volumetric measurements. There was a trend toward higher variability with decreasing radiation dose for both manual and semi-automated measurements of the longest diameter and volume (Table 1).

In regression models, quantitative measurements were only statistically different at 25% and 37.5% radiation dose levels for the semi-automated measurements of the longest diameter, as well as at 25% radiation dose level for semi-automated measurement of the volume (Fig. 2c). There was no significant difference in the variability of measurements of the longest

Fig. 2 (continued)



diameter and volume between the two commercially available semi-automated segmentation software (Table 2).

The variability related to changes in radiation dose was comparable to the inter- and intraobserver variability for both the manual and semi-automated measurements of the longest diameter ($p > 0.0008$ with Bonferroni correction). However, the variability related to changes in radiation dose for volume was significantly higher than that of inter- and intraobserver variability at 50% and 100% dose levels ($p < 0.0008$ with Bonferroni correction) (Table 3).

The mean observers' confidence for the accuracy of manual and semi-automated measurements of liver lesions at 100% dose level was 88.1 ± 8.7 (range 65–100) and 85.8 ± 8.0 (range 40–100), respectively. The regression model showed observers' confidence for the accuracy of the measurements was significantly worse at and below a $\leq 50\%$ radiation dose level (approximately $\text{CTDI}_{\text{vol}} = 4.80 \pm 1.53$ mGy), compared to the 100% dose level ($p \leq 0.003$) (Table 4 and Fig. 3).

The ICCs of different radiologists were excellent across all radiation dose levels for both manual (ICC 0.958 [95% CI 0.950–0.965]) and semi-automated measurements (ICC 0.963 [95% CI 0.955–0.969]) of the longest diameter, as well as good for the semi-automated measurement of volume (ICC 0.870 [95% CI 0.848–0.891], respectively) (Fig. 4).

Discussion

As the liver is the most common site of metastatic spread of colorectal cancer [26], there is a high demand for reliable, comparable and precise measurements of lesion size

for tumor staging and therapy response evaluation. Our results demonstrated that the variability related to radiation dose for both linear and volumetric measurements of the longest diameter was as high as the inter- and intraobserver variability. Given the worldwide trend in decreasing effective radiation dose [27], our preliminary results suggest that caution should be used when performing CT scans below 50% of a standard radiation dose (approximately $\text{CTDI}_{\text{vol}} = 4.80 \pm 1.53$ mGy in our study) due to its impact on measurements of lesion size may potentially result in misclassification of the overall response assessment in the liver [2]. The absolute difference and absolute percentage difference between 25 and 100% dose level of manual and semi-automated measurements of the longest diameter were as high as 1.4 cm (range 0.7–2.7 cm), 6.1% (range 3.0–11.5%) and 1.8 cm (range 0.7–4.0 cm), 7.0% (range 3.4–14.2%), respectively.

In our study, there was a trend toward higher measurement variability and decreased observer's confidence with decreasing radiation dose levels, for both manual and semi-automated measurements of the lesion's longest diameter and volume. Our results differ from a previous clinical study showing no impact of radiation dose on semi-automated volumetric measurements of lung nodules [20]. One possible explanation for this inconsistency may be the difference in lesion-to-background tissue contrast when comparing lesion measurements in the lung and liver. In contrast to the high lesion-to-background difference of lung nodules, the lesion-to-liver attenuation difference can be as low as 5–20 HU for colorectal liver metastases [28, 29]. In this context, an increase in noise with reduction in radiation dose may have

Table 1 Impact of radiation dose on measurements of the longest diameter and volume of liver metastasis

Predictors	Manual diameter			Semi-automated diameter			Semi-automated volume		
	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>
25%	1.01	1.00–1.03	0.085	1.05	1.03–1.07	<0.001	1.14	1.07–1.21	<0.001
37.5%	1.01	1.00–1.03	0.076	1.02	1.01–1.04	0.003	1.03	0.97–1.09	0.392
50%	1.00	0.99–1.02	0.666	1.01	1.00–1.03	0.124	1.03	0.97–1.10	0.357
62.5%	1.01	1.00–1.02	0.200	1.01	1.00–1.03	0.137	1.04	0.98–1.11	0.180
75%	1.00	0.99–1.02	0.661	1.00	0.99–1.02	0.679	1.01	0.95–1.08	0.748
87.5%	1.00	0.99–1.02	0.486	1.01	0.99–1.02	0.476	0.99	0.93–1.05	0.674

The results above were exponentiated from a linear mixed-effects regression model, with random intercepts for each reader and each lesion per patient. A log transformation was applied to the measurements

Bold *p* values are statistical significance

profound detrimental effects on liver lesion detection and measurement reproducibility.

Our results showed that both inter- and intraobserver variability of manual measurements tended to be higher than that of semi-automated measurements of the longest diameter, for all radiation dose levels. Our results are in accordance with previous studies in lung cancer [9], lymph node [30], and liver metastases [31]. It might be due to the fact that the slice position of each target lesion was pre-selected by one senior observer in our study. As a result, the key reason for the variability of semi-automated segmentation was the lesion’s attenuation. Human observers might be affected by other parameters as well, such as window level and width, image noise, lesion shapes, and internal texture.

The variabilities of semi-automated measurement of volume and the longest diameter have been reported inconsistently in the literature, with some studies showing increased variability [31, 32], some studies showing decreased variability [5, 30, 33, 34], and some studies showing comparable variability of volumetric measurements compared with linear measurements of the longest diameter [16, 35]. One possible explanation might be the difference in calculation of the variance, software for segmentation, selection of target lesions and scanning parameters among different studies. One previous phantom study [36] showed that lesion size, tissue contrast, imaging slice thickness, dose, and scanner were factors substantially influencing volume estimation of liver lesions. Our results further provide additional evidence regarding the importance of radiation dose as an additional and independent source of variability during manual and semi-automated measurements of metastatic liver lesions.

Some limitations of our study merit consideration. First, our clinical findings reflect experience from a relatively small patient cohort (*n* = 23). However, this is the first study comparing manual and semi-automated measurements across a broad range of dose levels (25–100%) within the same patient and at the same imaging time point in the liver. Second, due to the clinical nature of our study, we do not have a ground truth of liver lesion measurements leaving the true measurement error unknown. Third, our results are limited to a single dual-energy CT vendor and reconstruction setting, additional studies are needed to assess the impact of other reconstruction settings in lesion measurement variability.

In conclusion, radiation dose is a non-negligible source of variability for both manual and semi-automated measurements of metastatic liver lesions, being comparable to inter- and intraobserver variability in our study. Caution should be warranted in reducing the radiation dose level below 50% of a conventional CT protocol due to a significant increase in lesion measurement variability and potentially detrimental impact on the assessment of lesion response in the liver.

Table 2 Variability related to radiation dose between two semi-automated measurement tools of the longest diameter and volume of liver metastasis

Dose (%)	Longest Diameter			Volume		
	Vendor A	Vendor B	<i>p</i> value	Vendor A	Vendor B	<i>p</i> value
25	6.5% (3.2, 14.1)	7.5% (3.9, 14.4)	0.423	16.8% (7.5, 40.0)	17.7% (6.4, 41.2)	0.929
38	6.2% (2.8, 10.8)	7.9% (3.2, 14.0)	0.161	17.3% (6.0, 33.4)	15.1% (5.4, 33.8)	0.857
50	5.1% (1.9, 10.9)	6.0% (3.1, 12.3)	0.080	16.5% (5.7, 38.9)	14.3% (7.4, 26.5)	0.835
63	4.2% (1.8, 8.6)	5.8% (2.3, 11.0)	0.068	11.7% (5.9, 21.2)	14.0% (5.1, 24.5)	0.637
75	5.0% (1.7, 9.7)	5.4% (2.5, 9.9)	0.208	13.8% (7.7, 27.3)	12.0% (5.2, 24.9)	0.153
88	3.8% (1.8, 6.6)	5.4% (2.7, 11.2)	0.009	11.7% (5.6, 20.0)	9.8% (3.6, 21.7)	0.826

Data are median with interquartile range in parentheses. Vendor A = MintLesion; Vendor B = Syngo.Via

Table 3 *P* values of pairwise comparison among different types of variability

Dose level	Manual diameter			Semi-automated diameter			Semi-automated volume		
	inter versus intra	inter versus dose	intra versus dose	inter versus intra	inter versus dose	intra versus dose	inter versus intra	inter versus dose	intra versus dose
25%	0.6206	0.2476	0.6423	0.0891	0.2127	0.5061	0.1211	0.0083	0.0131
37.5%	0.8229	0.7988	0.6601	0.2698	0.0506	0.5255	0.0619	0.0964	0.7641
50%	0.2486	0.8074	0.3318	0.0496	0.0271	0.9262	0.7048	0.0001	0.0092
62.5%	0.0014	0.2546	0.0248	0.1407	0.6102	0.0499	0.9204	0.9678	0.8927
75%	0.8527	0.9681	0.8756	0.1582	0.0324	0.6406	0.0069	0.8804	0.0073
87.5%	0.8662	0.7862	0.9557	0.3584	0.5624	0.1431	0.3588	0.1360	0.6585
100%	0.5063	0.1122	0.0556	0.2704	0.3640	0.7470	0.6689	0.0007	0.0001

inter=interobserver variability; intra=intraobserver variability; dose=variability related to radiation dose. Bonferroni correction: $0.05/63=0.0008$ is used for pairwise comparison with the differences among different types of variability. *P* values are from Wilcoxon rank-sum tests comparing absolute percent differences from reference

Bold *p* values are statistical significance

Table 4 Impact of radiation dose on observers’ measurement confidence

Predictors (%)	Estimates	Confidence interval	<i>p</i> value
25	−5.60	−7.65 to −3.55	<0.001
37.5	−3.12	−5.17 to −1.07	0.003
50	−3.31	−5.36 to −1.26	0.002
62.5	−0.89	−2.94 to 1.17	0.397
75	−1.09	−3.14 to 0.96	0.296
87.5	0.26	−1.79 to 2.31	0.803

The effect of radiation dose on observers’ confidence was modeled using a linear mixed-effects regression model, with random intercepts for each reader and each lesion per patient

Bold *p* values are statistical significance

Fig. 3 Box plot shows mean measurement confidence of all observers across all dose levels. Observers' measurement confidence is significantly worse below a $\leq 50\%$ radiation dose level, compared to the 100% dose level ($p \leq 0.003$)

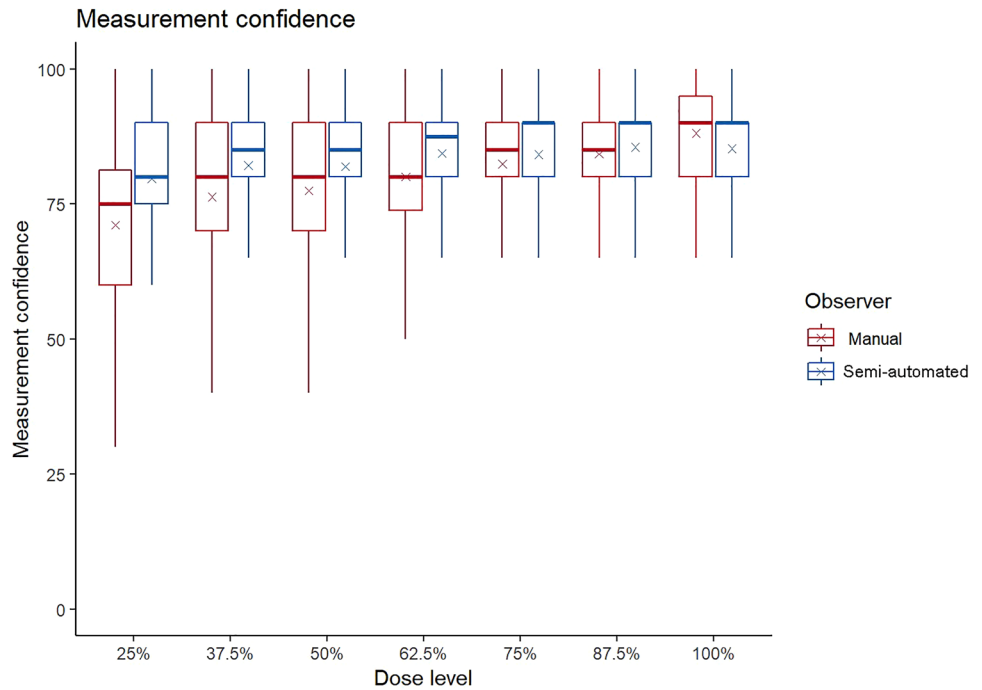
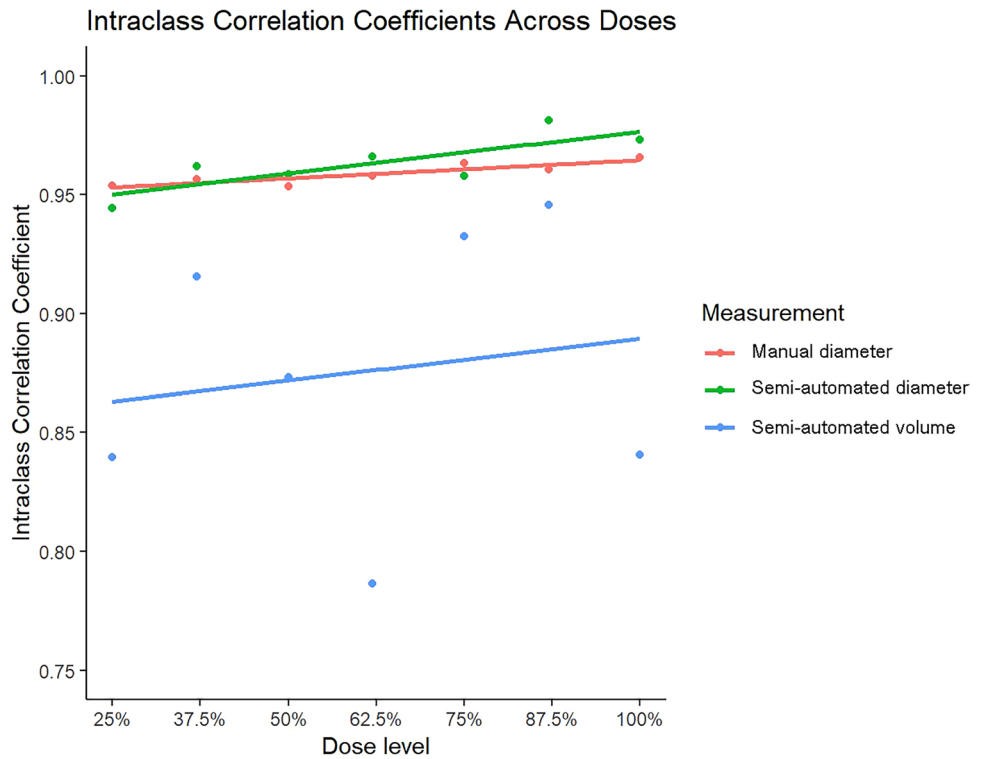


Fig. 4 Linear line graph of intraclass correlation coefficients (ICC) for all dose levels shows good to excellent consistency of both manual and volume measurements of liver lesions. The lines represent linear models of trends in ICC across dose levels



Author contributions MM, DM, YD: Study concepts. MM, DM, JS, ES: Study design. MM, DM, FV, FG: Data acquisition. MM, DM, ES: Quality control of data and algorithms. MM, DM, FV, FG, HVW: Data

analysis and interpretation. HVW: Statistical analysis. MM, DM, YD: Manuscript preparation. YD, MM, DM, FV, FG, HVW, HB, BNP, JS, ES, JCR, RCN: Manuscript editing. YD, MM, DM, FV, FG, HVW, HB, BNP, JS, ES, JCR, RCN: Manuscript review.

Funding None.

Data availability The corresponding author has full control of the data.

Compliance with ethical standards

Conflict of interest One author of the study (J.C.R.G.) is an employee of Siemens Healthineers, two authors (D.M. and M.M.) received research support (provision of software tools used in this study) from Siemens Healthineers and MintLesion. All other authors are not employees of or consultants for the industry and had control of any data or information that might present a conflict of interest. No conflict of interest exists in the submission of this manuscript, and manuscript is approved by all authors for publication. All the authors listed have approved the manuscript that is enclosed.

Ethics approval This study was approved by the institutional review board of Duke University and a waiver of written informed consent was obtained.

References

1. Therasse P, Arbutk SG, Eisenhauer EA, et al. (2000) 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* 92:205–216.
2. Eisenhauer EA, Therasse P, Bogaerts J, et al. (2009) New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 45:228–247.
3. Sargent DJ, Rubinstein L, Schwartz L, et al. (2009) Validation of novel imaging methodologies for use as cancer clinical trial endpoints. *Eur J Cancer* 45:290–299.
4. Prasad SR, Jhaveri KS, Saini S, Hahn PF, Halpern EF, Sumner JE. (2002) CT tumor measurement for therapeutic response assessment: comparison of unidimensional, bidimensional, and volumetric techniques initial observations. *Radiology* 225:416–419.
5. Mantatzis M, Kakolyris S, Amarantidis K, Karayiannakis A, Prassopoulos P. (2009) Treatment response classification of liver metastatic disease evaluated on imaging. Are RECIST unidimensional measurements accurate? *Eur Radiol* 19:1809–1816.
6. Suzuki C, Torkzad MR, Jacobsson H, et al. (2010) Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. *Acta Oncol* 49:509–514.
7. Bauknecht HC, Romano VC, Rogalla P, et al. (2010) Intra- and interobserver variability of linear and volumetric measurements of brain metastases using contrast-enhanced magnetic resonance imaging. *Invest Radiol* 45:49–56.
8. Bonekamp D, Bonekamp S, Halappa VG, et al. (2014) Interobserver agreement of semi-automated and manual measurements of functional MRI metrics of treatment response in hepatocellular carcinoma. *Eur J Radiol* 83:487–496.
9. Dinkel J, Khalilzadeh O, Hintze C, et al. (2013) Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer. *Lung Cancer* 82:76–82.
10. Erasmus JJ, Gladish GW, Broemeling L, et al. (2003) Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J Clin Oncol* 21:2574–2582.
11. Karademir I, Ward E, Peng Y, et al. (2016) Measurements of Hepatic Metastasis on MR Imaging: Assessment of Interobserver and Intersequence Variability. *Acad Radiol* 23:132–143.
12. Krajewski KM, Nishino M, Franchetti Y, Ramaiya NH, Van den Abbeele AD, Choueiri TK. (2014) Intraobserver and interobserver variability in computed tomography size and attenuation measurements in patients with renal cell carcinoma receiving antiangiogenic therapy: implications for alternative response criteria. *Cancer-Am Cancer Soc* 120:711–721.
13. McErlean A, Panicek DM, Zabor EC, et al. (2013) Intra- and interobserver variability in CT measurements in oncology. *Radiology* 269:451–459.
14. Kuhl CK, Alparslan Y, Schmoe J, et al. (2019) Validity of RECIST Version 1.1 for Response Assessment in Metastatic Cancer: A Prospective, Multireader Study. *Radiology* 290:349–356.
15. Zhao B, Schwartz LH, Moskowitz CS, et al. (2005) Pulmonary metastases: effect of CT section thickness on measurement—initial experience. *Radiology* 234:934–939.
16. Zhao B, Tan Y, Bell DJ, et al. (2013) Exploring intra- and interreader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals. *Eur J Radiol* 82:959–968.
17. Sakai N, Yabuuchi H, Kondo M, et al. (2015) Volumetric measurement of artificial pure ground-glass nodules at low-dose CT: Comparisons between hybrid iterative reconstruction and filtered back projection. *Eur J Radiol* 84:2654–2662.
18. Mileto A, Nelson RC, Larson DG, et al. (2017) Variability in Radiation Dose From Repeat Identical CT Examinations: Longitudinal Analysis of 2851 Patients Undergoing 12,635 Thoracoabdominal CT Scans in an Academic Health System. *AJR Am J Roentgenol* 208:1285–1296.
19. Jaffe TA, Yoshizumi TT, Toncheva G, et al. (2009) Radiation dose for body CT protocols: variability of scanners at one institution. *AJR Am J Roentgenol* 193:1141–1147.
20. Young S, Kim HJ, Ko MM, Ko WW, Flores C, McNitt-Gray MF. (2015) Variability in CT lung-nodule volumetry: Effects of dose reduction and reconstruction methods. *Med Phys* 42:2679–2689.
21. Solomon J, Zhang Y, Marin D, Samei E. (2016) SU-G-206-13: Validating Dose Split: A Method to Image the Same Patient at Multiple Doses with a Single CT Acquisition. *Med Phys* 43 (6Part25):36–42.
22. Menzel H, Schibila H, Teunen D. (2000) European Guidelines on Quality Criteria for Computed Tomography. Luxembourg: European Commission Publication.
23. Sica GT. Bias in research studies. (2006) *Radiology* 238:780–789.
24. Matthias Gamer, Jim Lemon and Ian Fellows Puspendra Singh. irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84. <https://www.r-project.org>. Published Jan 26, 2019. Accessed Mar 1, 2019.
25. Koo TK, Li MY. (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 15:155–163.
26. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology. Colon Cancer, Version 3.2019. Published Sept 26, 2019. https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf.
27. McCollough CH, Chen GH, Kalender W, et al. (2012) Achieving routine submillisievert CT scanning: report from the summit on management of radiation dose in CT. *Radiology* 264:567–580.
28. Kanal KM, Chung JH, Wang J, et al. (2011) Image noise and liver lesion detection with MDCT: a phantom study. *AJR Am J Roentgenol* 197:437–441.
29. Mileto A, Zamora DA, Alessio AM, et al. (2018) CT Detectability of Small Low-Contrast Hypoattenuating Focal Lesions: Iterative

- Reconstructions versus Filtered Back Projection. *Radiology* 289:443-454.
30. Buerke B, Puesken M, Muter S, et al.(2010)Measurement accuracy and reproducibility of semiautomated metric and volumetric lymph node analysis in MDCT. *AJR Am J Roentgenol* 195:979-985.
 31. van Kessel CS, van Leeuwen MS, Witteveen PO, Kwee TC, Verkooijen HM, van Hillegersberg R.(2012)Semi-automatic software increases CT measurement accuracy but not response classification of colorectal liver metastases after chemotherapy. *Eur J Radiol* 81:2543-2549.
 32. Zhao B, James LP, Moskowitz CS, et al.(2009)Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* 252:263-272.
 33. Keil S, Plumhans C, Behrendt FF, et al.(2009)Semi-automated quantification of hepatic lesions in a phantom. *Invest Radiol* 44:82-88.
 34. Fabel M, von Tengg-Kobligk H, Giesel FL, et al.(2008)Semi-automated volumetric analysis of lymph node metastases in patients with malignant melanoma stage III/IV--a feasibility study. *Eur Radiol* 18:1114-1122.
 35. Kalkmann J, Ladd SC, de Greiff A, Forsting M, Stattaus J.(2010) Suitability of semi-automated tumor response assessment of liver metastases using a dedicated software package. *Rofo* 182:581-588.
 36. Li Q, Liang Y, Huang Q, et al.(2016)Volumetry of low-contrast liver lesions with CT: Investigation of estimation uncertainties in a phantom study. *Med Phys* 43:6608.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.