


BMJ Open Retrospective comparison of approaches to evaluating inter-observer variability in CT tumour measurements in an academic health centre

MinJae Woo,¹ Moonseong Heo,¹ A Michael Devane ,² Steven C Lowe,² Ronald W Gimbel ¹

To cite: Woo M, Heo M, Devane AM, *et al*. Retrospective comparison of approaches to evaluating inter-observer variability in CT tumour measurements in an academic health centre. *BMJ Open* 2020;**10**:e040096. doi:10.1136/bmjopen-2020-040096

► Prepublication history and additional materials for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-040096>).

Received 05 May 2020

Revised 13 October 2020

Accepted 21 October 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Public Health Sciences, Clemson University, Clemson, South Carolina, USA

²Radiology, Prisma Health Upstate, Greenville, South Carolina, USA

Correspondence to

Dr Ronald W Gimbel;
rgimbel@clemson.edu

ABSTRACT

Background A growing number of research studies have reported inter-observer variability in sizes of tumours measured from CT scans. It remains unclear whether the conventional statistical measures correctly evaluate the CT measurement consistency for optimal treatment management and decision-making. We compared and evaluated the existing measures for evaluating inter-observer variability in CT measurement of cancer lesions.

Methods 13 board-certified radiologists repeatedly reviewed 10 CT image sets of lung lesions and hepatic metastases selected through a randomisation process. A total of 130 measurements under RECIST 1.1 (Response Evaluation Criteria in Solid Tumors) guidelines were collected for the demonstration. Intraclass correlation coefficient (ICC), Bland-Altman plotting and outlier counting methods were selected for the comparison. The each selected measure was used to evaluate three cases with observed, increased and decreased inter-observer variability.

Results The ICC score yielded a weak detection when evaluating different levels of the inter-observer variability among radiologists (increased: 0.912; observed: 0.962; decreased: 0.990). The outlier counting method using Bland-Altman plotting with 2SD yielded no detection at all with its number of outliers unchanging regardless of level of inter-observer variability. Outlier counting based on domain knowledge was more sensitised to different levels of the inter-observer variability compared with the conventional measures (increased: 0.756; observed: 0.923; improved: 1.000). Visualisation of pairwise Bland-Altman bias was also sensitised to the inter-observer variability with its pattern rapidly changing in response to different levels of the inter-observer variability.

Conclusions Conventional measures may yield weak or no detection when evaluating different levels of the inter-observer variability among radiologists. We observed that the outlier counting based on domain knowledge was sensitised to the inter-observer variability in CT measurement of cancer lesions. Our study demonstrated that, under certain circumstances, the use of standard statistical correlation coefficients may be misleading and result in a sense of false security related to the consistency of measurement for optimal treatment management and decision-making.

Strengths and limitations of this study

- While several conventional statistical measures are frequently used to evaluate inter-observer variability in radiological measurement, very few comparative studies have been performed to quantify the relative merits of the measures.
- The study demonstrated there is no evidence to support the use of statistical correlation coefficient for the assessment of inter-observer CT measurement variability.
- This is a retrospective study conducted in a single academic health centre.
- Another limitation may be the measurements collected under a highly controlled environment where the radiologists were rarely interrupted throughout the data collection.

BACKGROUND

Clinical evaluation of cancer therapeutics is based on the assessment of change in tumour burden, which is an important surrogate marker reflecting the therapeutic efficacy of cancer treatments. A comprehensive evaluation of tumour burden often involves a series of measurements of multiple tumour diameters. Measurement accuracy and consistency are essential; a large inter-observer variability in measuring tumour size may interfere with precise assessment of cancer treatment response when serial measurements are performed by multiple radiologists. Some studies suggest there are radiologist-dependent factors (eg, preferred guideline, measurement technique, years of clinical experience) that may contribute variability in the anatomical measurements.^{1–6} A potentially heightened patient risk associated with the inter-observer variability may be present when a patient's repeat CT imaging is assigned to a radiologist different from the radiologist who originally measured the

tumour. As a result, clinical disagreement due to the variability between the radiologists may result in an unnecessary change in treatment management.

Predominant methods for evaluation of the inter-observer variability in radiological measurements typically include measures based on statistical correlation coefficient and Bland-Altman plot.^{2 7-14} Intraclass correlation coefficient (ICC) is a widely used reliability measure comparing the variability of different ratings by the same raters to the total variation across all ratings and all raters.¹⁵ This reliability measure can be used for test-retest, intra-rater and inter-rater reliability analyses when the rating scale is continuous or ordinal. The Bland-Altman plotting is another popular exploratory analysis approach for intra-rater and inter-rater reliability when two paired measurements use the same scale.¹⁶

While these measures serve as useful assessment instruments in many other fields,¹⁷⁻²⁰ their use in evaluating the variability in radiological measurements has not been adequately explored. There is a paucity of research investigating either the absolute or comparative effectiveness of these measures in evaluating inter-observer measurement variability among radiologists. Despite multiple statistical studies containing an explicit warning against the use of correlation-based measures and visualisation in some cases,^{15 21-25} it remains unclear whether the measures are sufficiently responsive to appropriately evaluate the inter-observer variability. Consequently, it is also not known whether these measures can be used for interventional studies aiming to reduce inter-observer variability in measurement.⁶ Previous studies on inter-observer variability in radiological measurement have reported correlation coefficient scores ranging from 0.860 to 0.999.^{2 7-11 14} From a radiologist's perspective, these numbers offer little clinical insight on level of the inter-observer variability other than the fact that the scores are very high. The question of how high score is small inter-observer variability is open for further investigation.

In this paper using cases with different levels of inter-observer measurement variability, we compare sensitivity and clinical usefulness of different evaluation measures for inter-observer variability in CT lesion measurements. Additionally, cases were assessed using these measures to offer a better clinical insight for the question of how high the scores should be to achieve clinically acceptable measurement variability in daily clinical practice.

METHODS

Our demonstration is based on three cases with increased, observed and decreased inter-observer measurement variability that were generated from real clinically observed data. Descriptions of how data were generated for each case are detailed below. The observed data set was acquired from a single-site, double-blinded, observational study, conducted in the Department of Radiology, Prisma Health System, located within the Southeast USA. The study was conducted between July 2017 to December 2017.

The Department of Radiology operates in an academic health centre but does not train radiology residents.

Collecting observed data

Data were collected from 13 board-certified radiologists who regularly read CT examinations of lung lesions and hepatic metastasis. Each of the five lung lesions and five hepatic metastases samples were randomly selected from the Picture Archiving and Communication System (PACS) following two primary criteria: (1) whether the lesions are measurable under the Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 guideline, and (2) whether the lesions are commonly encountered in clinical practice. See online supplemental material 1, which are the selected images. These CT images contained normal anatomy cephalad and caudal to the lesion of interest. Each CT image set did not contain any recommendations regarding measurement. The 13 radiologists independently reviewed the same 10 CT image sets, which resulted in a total of 130 measurements (13×10). Individual radiologists adjusted the window level according to their preferences, as they would in their clinical practice. According to RECIST 1.1 criteria, only the longest CT axis of a tumour image and its corresponding measurement were collected.

Creating cases with different levels of inter-observer variability

The original observed data were used to generate cases with increased, observed and decreased inter-observer measurement variability. The extent of variability classified as increased, observed or decreased does not indicate the absolute level of measurement variability; the classifications were used to indicate different cases with relatively high or relatively low inter-observer variability. The original observed data served as the data representing the case with observed inter-observer measurement variability.

We generated data representing the case with increased inter-observer variability by moving each measurement in the observed data away from the nearest peer measurements. Specifically, we inflated the inter-observer variability by increasing the deviation of each measurement from the corresponding median by 40% to create a case with evidently unacceptable measurement variability. Similarly, the deviation of each measurement from the corresponding median was decreased by 40% in the case with decreased inter-observer variability, [figure 1](#). The per cent differences between each measurement and the corresponding median were visualised using scatter plots for all CT image sets, [figure 2](#). The raw data for each case can be found in online supplemental material 2.

Description of selected measures for comparison

We selected evaluation measures based on ICC and Bland-Altman plot, which are commonly used for the assessment of intra-observer and inter-observer variability in CT measurement.^{2 7-14} While Bland-Altman plot is a

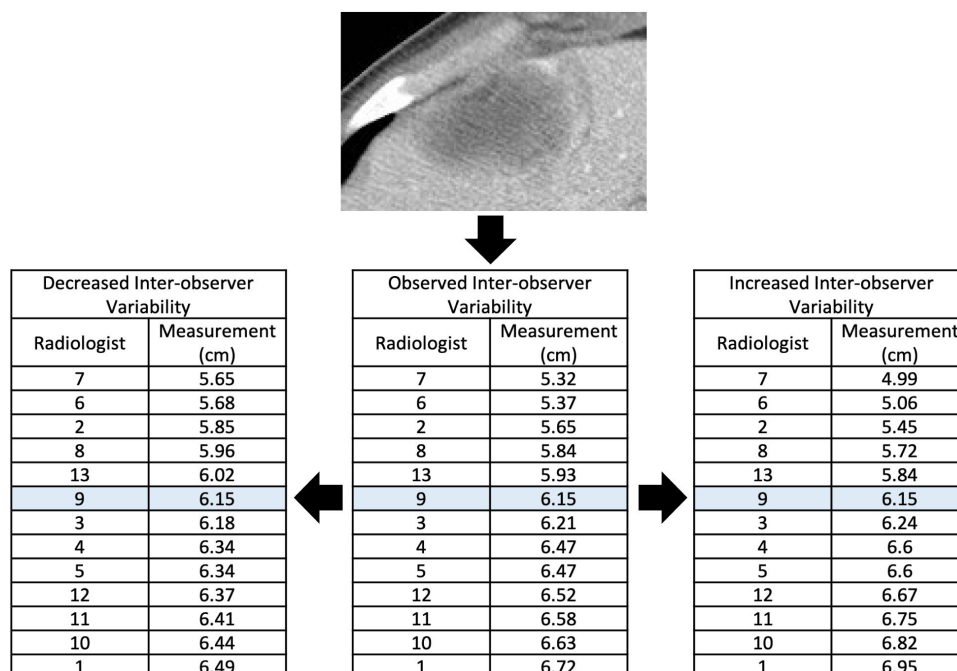


Figure 1 Example of increased and decrease inter-observer variability from observed data. To generate a case with increased inter-observer variability, the difference between each measurement and the median value was increased by 40% (right). The difference between each measurement and the median value was decreased by 40% in the case with decreased inter-observer variability (left).

graphical method rather than statistical measure, some well-respected studies used the plotting for tracking a number of outlier measurement differences outside the 2SD upper and lower limit of agreement (LOA).^{2 14 26} Accordingly, we quantified Bland-Altman plots using a number of data points exceeding the upper and lower LOA. The plotting compares two radiologists at a time; for each case, we performed a pairwise Bland-Altman analysis for all possible pairs within a group of radiologists and counted the total number of outliers from all pairs, online supplemental material 3. If the number of outliers from Bland-Altman plot is sensitised to the different levels of inter-observer variability, more outliers (ie, higher proportion of outlier measurement differences) would

be observed in the case with increased inter-observer variability.

In the clinical context, this pairwise approach explores how safely a patient can be transferred from one radiologist to another within a group of radiologists. If two radiologists reviewed the same set of CT cases but suggested measurements largely different from each other, there may be concerns associated with the patient transfer between the radiologists. Similarly, if two radiologists reviewed the same set of CT cases and suggested measurements similar to each other, the concerns associated with the patient transfer may be marginal. Having more pairs with fewer outlier measurement differences may imply

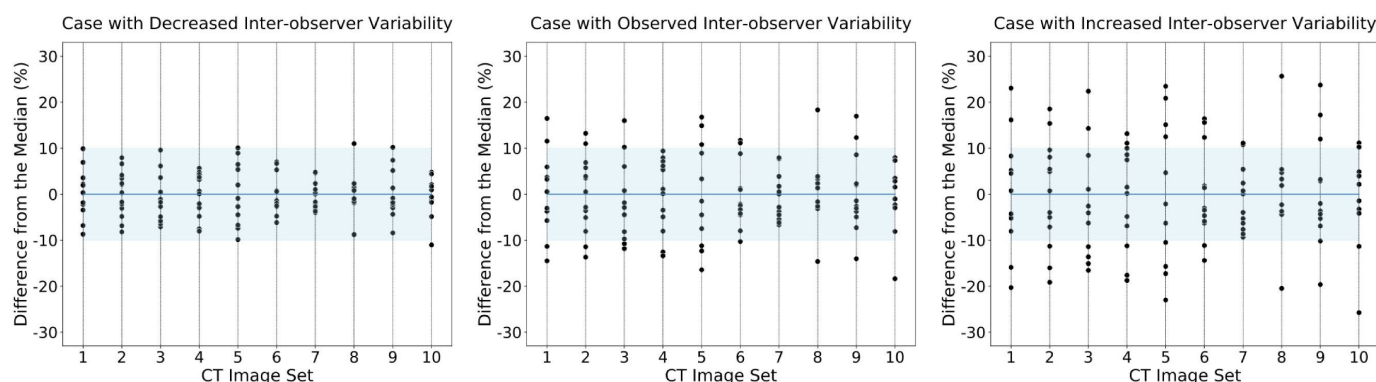


Figure 2 Visualisation of measurement distribution for each case. Each vertical line in the graphs represent different CT case and each point represent per cent difference between a measurement and the corresponding median value. The light blue area represents plus and minus 10% interval from the median value.

less concern for inter-observer variability when a patient is reviewed by multiple radiologists.

Statistical analysis

We compared three evaluation measures for the comparison: (1) ICC, (2) Bland-Altman plot with 2SD LOA and (3) Bland-Altman plot with 20% fixed LOA. As for estimations of ICC scores, a two-way random-effects model that characterises absolute agreement by incorporating both lesion-wise effect (target effect) and radiologist-wise effect (rater effect) was applied for both simulated and observed data.^{2 19 27 28} The ICC scores were estimated based on all 130 measurements for each case (increased, observed and decreased).

While Bland-Altman plot allows data to be analysed both as unit differences plot and as percentage differences plot,¹⁶ we used per cent difference plot as suggested by previous studies in the literature.^{2 14 28} Bland-Altman plot with 2SD LOA was quantified into score value by calculating proportion of data points within the upper and lower LOA.

Bland-Altman plot with 20% fixed limits was also quantified into score value to compare with ICC and standard Bland-Altman plot with 2SD limits. There have been several clinical studies using Bland-Altman plot with fixed limits of agreement evidenced by relevant domain knowledge.^{29 30} This essentially aligns with other studies that use clinical domain knowledge to define outliers.^{31–34} We fixed the maximum acceptable LOA to assess the measurement interchangeability between radiologists at 20% evidenced by clinical guidelines. The predominant guideline for cancer treatment response evaluation, RECIST 1.1, heavily depends on per cent difference in lesion diameter with a progression defined as a 20% increase in the sum of longest diameters.^{35 36} The absolute inter-radiologist difference already exceeding 20% in CT measurements may interfere with the application of the 20% criterion from the guideline when a patient is reviewed by different radiologists. Thus, the 20% measurement difference was

used as the fixed LOA for the Bland-Altman plot. In the context of radiological measurement, this means that outlier measurement difference is explicitly defined as measurement difference exceeding 20% when a pair of radiologists reviewing the same image.

Bland-Altman plot also allows identification of any systematic difference (mean difference in measurements) between two observers. For each case of inter-observer variability, the mean difference in measurements was calculated for all possible pairs ($n=78$) and visualised in a heat map, figure 3.

Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

RESULTS

Characteristics of CT image sets included in the study

Each CT image set included in the study consisted of multiple CT slices with an average of 7.6 images, table 1. The minimum and maximum size of the hepatic metastases ranged between 1.68 cm to 2.21 cm and 5.32 cm to 6.72 cm, respectively. The minimum and maximum size of lung lesions ranged between 1.27 cm to 1.68 cm and 3.69 cm to 5.02 cm, respectively. In the observed data, the largest lesion-wise per cent difference in measurements was realised in Hepatic Metastasis 5 with 33.1% difference between the minimum and maximum measurements. The smallest lesion-wise per cent difference in measurements was realised in Lung Lesion 2 with 14.5% difference between the minimum and maximum measurements.

Characteristics of cases with different levels of inter-observer variability

The graph visualisation of the data from each case suggested varying levels of inter-observer variability, figure 2. The visualisation of the original observed data

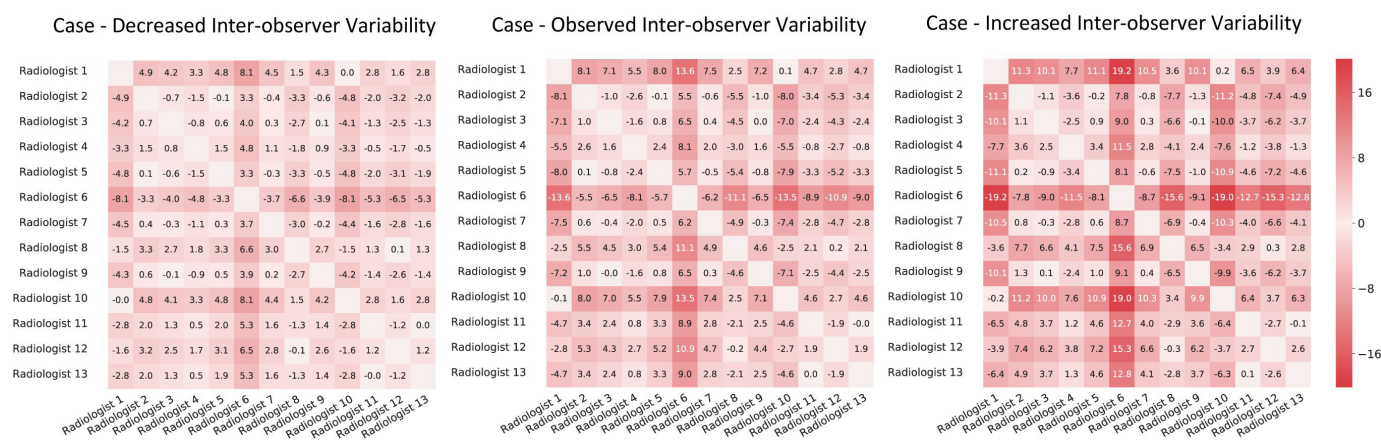


Figure 3 Visualisation of pairwise bias from Bland-Altman analysis. The systematic discrepancy (bias) was calculated using average per cent differences and presented in decimal format. Darker red colours represent larger per cent measurement differences. The positive values indicate that the radiologist on y-axis over-estimated compared with the radiologist on x-axis. The negative values indicate that the radiologist on y-axis under-estimated compared with the radiologist on x-axis.

Table 1 Descriptive statistics for the original observed data

CT image sets	Number of image slices	Median measurements (SD)	Range	Min–Max per cent difference
Hepatic metastasis 1	9	4.46 (0.38)	(3.81 to 5.19)	30.70%
Hepatic metastasis 2	5	2.68 (0.22)	(2.31 to 3.03)	27.00%
Hepatic metastasis 3	5	1.91 (0.18)	(1.68 to 2.21)	27.20%
Hepatic metastasis 4	13	6.14 (0.48)	(5.32 to 6.72)	23.30%
Hepatic metastasis 5	6	2.68 (0.29)	(2.24 to 3.13)	33.10%
Lung lesion 1	8	3.46 (0.24)	(3.10 to 3.86)	21.80%
Lung lesion 2	10	4.18 (0.23)	(3.90 to 4.51)	14.50%
Lung lesion 3	6	2.00 (0.17)	(1.71 to 2.37)	32.40%
Lung lesion 4	10	4.29 (0.36)	(3.69 to 5.02)	30.50%
Lung lesion 5	4	1.56 (0.11)	(1.27 to 1.68)	27.80%

Note: Average measurement and range are in centimetres (cm). SD denotes standard deviation. Min denotes minimum measurement for each lesion. Max denotes maximum measurement for each lesion. Per cent difference between minimum and maximum values was calculated using the following formula: difference (min, max)/average (min, max). Range consists of (minimum observed value to maximum observed value).

suggested a substantial inter-observer variability with 31 (23.8%) measurements outside the light blue area representing plus or minus 10% interval from the average measurement value for each case. Additionally, a lesion-wise effect on inter-observer variability was observed with relatively high measurement variation in some CT image sets. The visualisation of the case of decreased inter-observer variability illustrated a small number of measurements outside the threshold with 3 (2.3%) measurements locating outside the plus or minus 10% interval. With the decrease in the deviations of each measurement from the corresponding median, all measurements moved towards average and closer together as intended for demonstration. On the other hand, there was a relatively large number of measurements outside the threshold in the case of increased inter-observer variability with 50 (38.5%) measurements locating outside the plus or minus 10% interval. Also, it was observed that all measurements were not only shifted away from median, but also moved further away from each other as intended.

Visualisation of Bland-Altman analysis

The heat map visualisation of average per cent measurement difference (fixed bias) for all pairs of radiologists suggested varying levels of the difference across all pairs, figure 3. Some pairs of radiologists achieved a lower average per cent difference than others. In the heat map of the original observed data, the smallest systematic difference in measurement was observed in the pair of Radiologist 11 and Radiologist 13; they maintained an average of 0.03% difference in their measurements when reviewing the same set of CT images. The largest systematic measurement difference was observed in the pair of Radiologist 1 and Radiologist 6. The systematic difference in their measurements was 13.6% when reviewing the same set of CT images. It was observed that some radiologists attributed more to inter-observer variability

than others; Radiologist 1 and 10 generally overestimate lesion size compared with others while Radiologist 2 and 6 generally underestimated lesion size compared with others.

The heat map visualisation from the case of increased inter-observer variability showed the increased systematic measurement differences between any two radiologists compared with other cases. Similarly, the heat map visualisation from the case of decreased inter-observer variability showed the decreased systematic measurement differences compared with other cases. Overall, the cases with relatively high inter-observer variability tend to present the increased systematic measurement differences between any two radiologists as well as more pairs of radiologists with a systematic measurement difference close to 20% when reviewing the same CT image sets.

Comparison of the selected measures

The original observed data achieved the ICC score of 0.962. The ICC scores in the cases of increased and decreased inter-observer variability were 0.990 and 0.912, respectively. The per cent increase in the deviation of each measurement from the corresponding median has a perfect linear relationship with the ICC score ($R^2=1.00$), figure 4. However, the magnitude of association was extremely low; 10 per cent increase in the deviation was associated with 0.01 decrease in the ICC score. As a result, the graph representing a relationship between a per cent increase in the deviation and the corresponding ICC score presented a virtually flat slope, which implies that the score is extremely insensitive to the changes in deviations.

The original observed data achieved the standard Bland-Altman score of 0.937, which indicates 93.7% of data points within lower and upper LOA along with 6.3% outlier data points. The score based on standard Bland-Altman presented flat slope with its score unchanging

Comparison of Intraclass Correlation and Bland-Altman Outlier Score

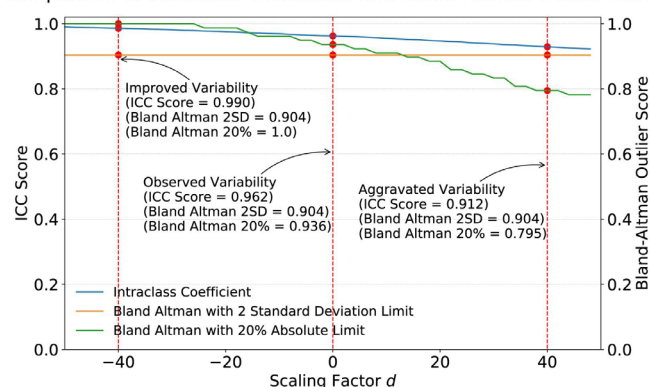


Figure 4 Responsiveness comparison of intraclass correlation coefficient (ICC) and Bland-Altman outlier scores. Scaling factor d represents per cent increase in the deviation of each measurement from the corresponding median. Horizontal axis corresponds to scaling factor d used to decrease or increase the inter-observer variability. Vertical axis represents ICC and Bland-Altman scores. Vertical dotted lines in red represent different data sets.

regardless of level of inter-observer variability (standard Bland-Altman score=0.937).

The presented Bland-Altman score with fixed limits was more responsive to the change in case than other measures. In the case with decreased inter-observer variability, all pairs were identified to have a per cent difference less than 20% when reviewing the same CT image sets (fixed-limit Bland-Altman score=1.0). The original observed data suggested Bland-Altman score with fixed limits of 0.923 with 92.3% of all possible pairwise measurements having a per cent difference less than 20%. In the case with increased inter-observer variability, 75.6% of measurements were identified to have a per cent difference less than 20% when reviewing the same CT image sets. The Bland-Altman score with fixed limits changed by 0.167 (0.756 to 0.923) between increased case and observed data, and 0.077 (0.923 to 1.000) between observed data and increased case, [figure 4](#).

DISCUSSION

The importance of consistent measurement of cancer lesions in CT scans has been well documented.^{10 35 36} We have performed an extensive simulation study using conventional evaluation measures and different cases with varying levels of inter-observer variability. Our study investigated precision of those measures and found that some measures are not sensitive enough to detect the difference between cases with clinically desirable and clinically unacceptable inter-observer variability in radiological measurement.

The previous studies by McErlean *et al* and Zhao *et al* used statistical correlation coefficients and standard Bland-Altman plot as primary measures and concluded that serial CT measurements can be safely performed by different radiologists.^{2 7} Our study indicated that the

correlation-based measures may fail to serve as a true indicator of inter-observer variability. When the observed data were analysed, the radiologists in our study achieved a high ICC score comparable to previous studies.^{2 13} However, as demonstrated above, a high ICC score does not always guarantee low inter-observer variability in the context of radiological measurement. Our analysis suggests that the statistical correlation-based measures may yield high scores regardless of level of the inter-observer variability among radiologists. Therefore, a group of radiologists who achieved a high ICC score within the group could fail to maintain clinically reasonable measurement consistency. For instance, an ICC score of 0.9 achieved by a group of readers is often considered to be excellent in many other fields.^{36 37} However, in the case of cancer treatment response evaluation, the ICC score of 0.9 may raise serious patient safety concerns with radiologists always having at least 10% average per cent difference in measurement to each other when reviewing the same CT image sets. In the presented case with increased inter-observer variability, the ICC score of 0.91 was still not high enough to achieve clinically acceptable inter-observer variability in CT measurement, as affirmed by the participating radiologists, online supplemental material 2. Despite the unrealistically high increase in the variability observed in the case with increased inter-observer variability, the ICC score failed to provide an adequate warning.

Another measure, outlier counts from standard Bland-Altman plotting with 2SD upper and lower LOA, presented no response to the varying levels of inter-observer variability in CT measurements. It was observed that its upper and lower limits increase proportionally to measurement variabilities, [figure 5](#). Our analysis suggested no evidence to support its use for the assessment of CT measurement variability or outlier detection.

While the standard Bland-Altman and ICC scores changed little across the different cases, the presented Bland-Altman score with 20% fixed limits rapidly changed between cases of increased, observed and decreased inter-observer variability. The presented score is also intuitive to interpret because of its self-descriptive nature; the decrease in the score from 0.923 to 0.756 means that the percentage of pairwise measurements having less than 20% difference has decreased from 92.3% to 75.6%. As documented, the predominant guideline for cancer treatment response evaluation defines a diameter increase of 20% as the cut-off for progression of cancer. If multiple pairs of measurements have 20% or higher measurement difference over the same CT image sets, this may interfere with the application of the 20% criterion from the guideline when a patient is reviewed by different radiologists. The Bland-Altman score with fixed limits demonstrated a potential to detect a decrease in the number of pairs having less than 20% measurement difference when reviewing the same image sets, which may better facilitate the application of guideline.

The Bland-Altman heat map of pairwise systematic discrepancy offered some useful insight on how the

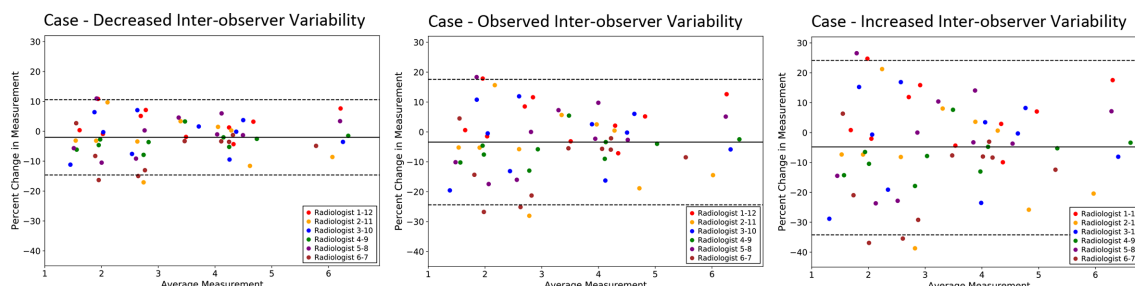


Figure 5 Standard Bland-Altman plotting for the selected pairs. The upper and lower limit of agreement (LOA) were calculated using 2SD. The dotted and solid lines represent LOA and mean difference, respectively. Different colours represent different radiologist pairs. While there were a total 78 possible pairs, the plotting included six selected pairs for visualisation purposes. The total number of outliers was unchanging across the different cases, regardless of the number of pairs in the plotting.

inter-observer variability can be addressed in interventional studies. The visualisation identified radiologists who largely under-measure or over-measure compared with their peers, which can be a potential target for intervention to reduce the variability. Risk associated with inter-observer variability is realised when a patient is referred from one radiologist to another or reviewed by different radiologists. The pairwise approach to visualise systematic discrepancy may also be useful in addressing the risk by identifying pair of radiologists whose measurements typically differ greatly from each other.

This was a retrospective study conducted in a single academic health centre. Future study may extend our approach to more measurements with various respond evaluation criteria used by radiologists from multiple institutions. A potential limitation of the study may result from the image selection process. Although the images were randomly selected from the health system PACS, the application of the selection criteria was performed by one senior radiologist. A selection criterion was whether or not images are commonly encountered in daily clinical practice, which may have introduced a bias in the image selection. Another limitation is that the measurements were collected under a highly controlled environment where the radiologists were rarely interrupted throughout the data collection. It is commonly believed that in real-world clinical practice, one's actual performance may be negatively affected by a heavy workload or various types of interruptions. Lastly, future studies are warranted to explore other existing evaluation approaches. For example, although the reliability of the estimated regression line depends on the sample size, the homoscedasticity and normality of the distribution of the differences, the regression of the mean on the difference could reveal whether the extend of disagreement depends on the mean of two measurements.

CONCLUSIONS

Conventional measures may yield weak or no detection when evaluating different levels of the inter-observer variability among radiologists. We observed that the outlier counting based on domain knowledge was sensitised to the inter-observer variability in CT measurement of cancer lesions. Our study demonstrated that, under certain circumstances, the use of standard statistical

correlation coefficients may be misleading and result in a sense of false security related to the consistency of measurement. A visualisation based on pairwise approach to identify systematic discrepancy may serve as a useful and practical tool for future efforts to reduce the inter-observer variability in radiological measurement.

Twitter A Michael Devane @amdevane

Acknowledgements The authors acknowledge and appreciate the participating radiologists of the Prisma Health System. We also acknowledge and appreciate the logistics and regulatory support of Karen Edwards, MS of the Department of Public Health Sciences, Clemson University.

Contributors MW designed the study, analysed the data and developed the manuscript. MH made substantial contributions to the data analysis and critical revisions. SCL and AMD acted as Clinical Investigators and contributed substantially to study development and clinical data interpretation. RWG served as co-Principal Investigator and supervised preparation, conduct and administration of the study. All authors developed, reviewed and approved the manuscript.

Funding This work was supported by supported by Health Science Center, Prisma Health, Greenville, South Carolina (Grant number: Pro00065670).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Institutional Review Board, Prisma Health System, Greenville, South Carolina.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. All data relevant to the study are included in the manuscript and uploaded in Supplementary Material 2.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

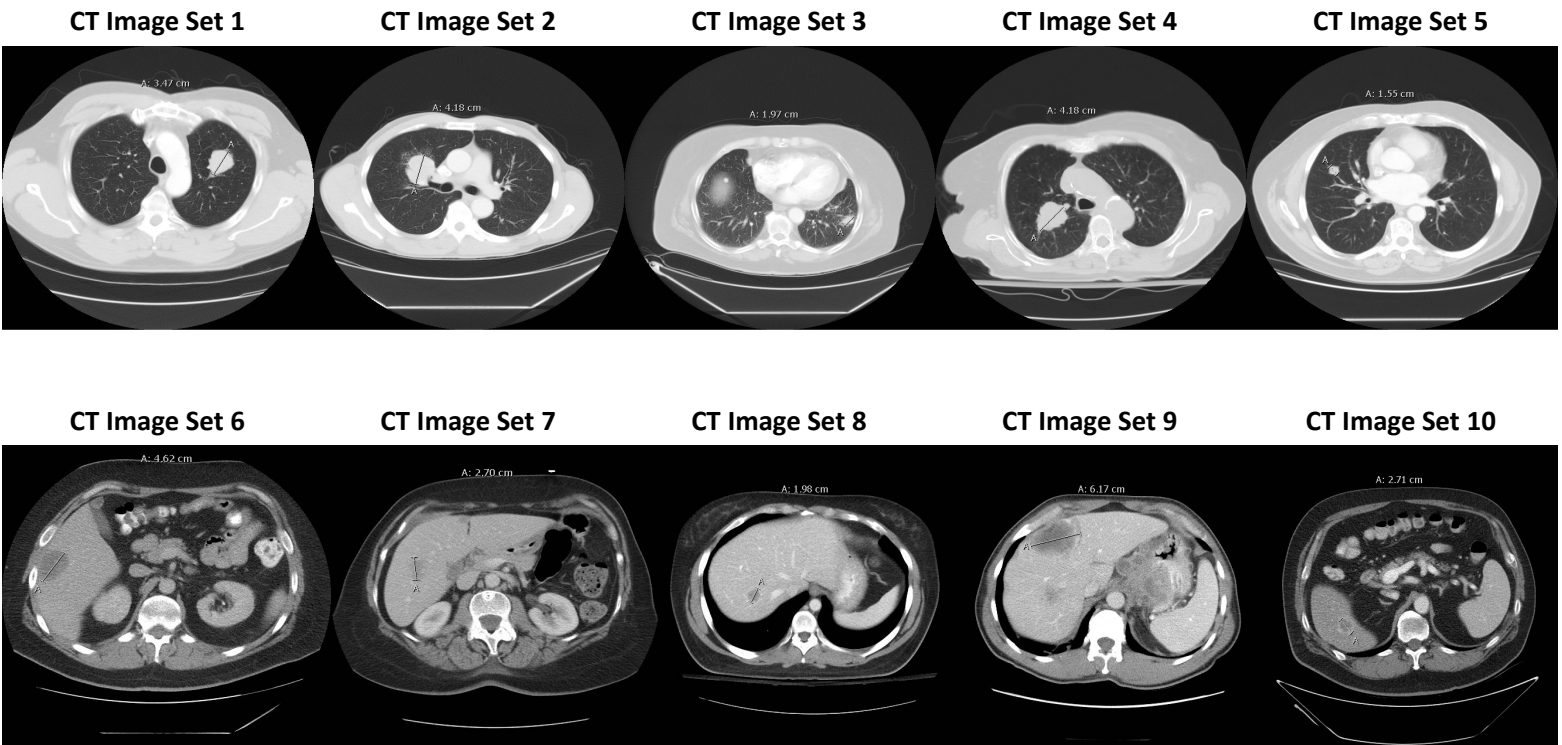
A Michael Devane <http://orcid.org/0000-0003-2497-8882>
Ronald W Gimbel <http://orcid.org/0000-0001-8185-4013>

REFERENCES

- 1 Jiang B, Zhou D, Sun Y, *et al.* Systematic analysis of measurement variability in lung cancer with multidetector computed tomography. *Ann Thorac Med* 2017;12:95–100.
- 2 McErlean A, Panicek DM, Zabor EC, *et al.* Intra- and interobserver variability in CT measurements in oncology. *Radiology* 2013;269:451–9.
- 3 Oxnard GR, Zhao B, Sima CS, *et al.* Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. *J Clin Oncol* 2011;29:3114–9.
- 4 Singh S, Maxwell J, Baker JA, *et al.* Computer-aided classification of breast masses: performance and interobserver variability of expert radiologists versus residents. *Radiology* 2011;258:73–80.
- 5 Thiesse P, Ollivier L, Di Stefano-Louineau D, *et al.* Response rate accuracy in oncology trials: reasons for interobserver variability. Groupe Français d'Immunothérapie de la Fédération Nationale des centres de Lutte Contre Le cancer. *J Clin Oncol* 1997;15:3507–14.
- 6 Woo M, Lowe SC, Devane AM, *et al.* Intervention to reduce interobserver variability in computed tomographic measurement of cancer lesions among experienced radiologists. *Curr Probl Diagn Radiol* 2020.
- 7 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;252:263–72.
- 8 Wormanns D, Diederich S, Lentschig MG, *et al.* Spiral CT of pulmonary nodules: interobserver variation in assessment of lesion size. *Eur Radiol* 2000;10:710–3.
- 9 Tyng CJ, Chojniak R, Pinto PNV, *et al.* Conformal radiotherapy for lung cancer: interobservers' variability in the definition of gross tumor volume between radiologists and radiotherapists. *Radiat Oncol* 2009;4:28.
- 10 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 Am J Roentgenol* 2010;195:W221–8.
- 11 Chung MS, Cheng KL, Choi YJ, *et al.* Interobserver reproducibility of cervical lymph node measurements at CT in patients with head and neck squamous cell carcinoma. *Clin Radiol* 2016;71:1226–32.
- 12 Cornelis FH, Martin M, Saut O, *et al.* Precision of manual two-dimensional segmentations of lung and liver metastases and its impact on tumour response assessment using RECIST 1.1. *Eur Radiol Exp* 2017;1:16.
- 13 Dinkel J, Khalilzadeh O, Hintze C, *et al.* Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer. *Lung Cancer* 2013;82:76–82.
- 14 Krajewski KM, Nishino M, Franchetti Y, *et al.* 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* 2014;120:711–21.
- 15 Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation - A discussion and demonstration of basic features. *PLoS One* 2019;14:e0219854.
- 16 Giavarina D. Understanding Bland Altman analysis. *Biochem Med* 2015;25:141–51.
- 17 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
- 18 Hemphill JF. Interpreting the magnitudes of correlation coefficients. *Am Psychol* 2003;58:78–9.
- 19 Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012;24:69–71.
- 20 Wolak ME, Fairbairn DJ, Paulsen YR. Guidelines for estimating repeatability. *Methods Ecol Evol* 2012;3:129–37.
- 21 Bobak CA, Barr PJ, O'Malley AJ. Estimation of an inter-rater intraclass correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Med Res Methodol* 2018;18:93.
- 22 Shoukri MM, Donner A. Efficiency considerations in the analysis of inter-observer agreement. *Biostatistics* 2001;2:323–36.
- 23 Weinberg R, Patel YC. Simulated intraclass correlation coefficients and their z transforms. *J Stat Comput Simul* 1981;13:13–26.
- 24 Ponzoni RW, James JW. Possible biases in heritability estimates from intraclass correlation. *Theor Appl Genet* 1978;53:25–7.
- 25 Ludbrook J. Confidence in altman-bland plots: a critical review of the method of differences. *Clin Exp Pharmacol Physiol* 2010;37:143–9.
- 26 Faria SL, Faria OP, Cardeal MDA, *et al.* Validation study of multi-frequency bioelectrical impedance with dual-energy X-ray absorptiometry among obese patients. *Obes Surg* 2014;24:1476–80.
- 27 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
- 28 Gutin B, Litaker M, Islam S, *et al.* Body-composition measurement in 9–11-year-old children by dual-energy X-ray absorptiometry, skinfold-thickness measurements, and bioimpedance analysis. *Am J Clin Nutr* 1996;63:287–92.
- 29 Bogui P, Balayssac-Siransy E, Connes P, *et al.* The PhysioFlow thoracic impedancemeter is not valid for the measurements of cardiac hemodynamic parameters in chronic anemic patients. *PLoS One* 2013;8:e79086–e86.
- 30 Vent-Schmidt J, Waltz X, Pichon A, *et al.* Indirect viscosimetric method is less accurate than ektacytometry for the measurement of red blood cell deformability. *Clin Hemorheol Microcirc* 2015;59:115–21.
- 31 Schold JD, Miller CM, Henry ML, *et al.* Evaluation of flagging criteria of United States kidney transplant center performance: how to best define outliers? *Transplantation* 2017;101:1373–80.
- 32 Bergamin O, Anderson SC, Kardon RH. An objective method to define outlier optical coherence tomograms and repeatability of retinal nerve fibre layer measurements. *Acta Ophthalmol Scand* 2004;82:535–43.
- 33 Montalbano A, Quiñonez RA, Hall M, *et al.* Achievable benchmarks of care for pediatric readmissions. *J Hosp Med* 2019;14:534–50.
- 34 Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics* 2006;7:123.
- 35 Nishino M, Jagannathan JP, Ramaiya NH, *et al.* Revised RECIST guideline version 1.1: what oncologists want to know and what radiologists need to know. *AJR Am J Roentgenol* 2010;195:281–9.
- 36 Schwartz LH, Litière S, de Vries E, *et al.* RECIST 1.1-Update and clarification: from the RECIST Committee. *Eur J Cancer* 2016;62:132–7.
- 37 Gellhorn AC, Carlson MJ. Inter-rater, intra-rater, and inter-machine reliability of quantitative ultrasound measurements of the Patellar tendon. *Ultrasound Med Biol* 2013;39:791–6.

Supplementary 1.

Examples of measurement for all CT image sets used in this study.



Supplementary 2.

To generate new data M'_{ik} for the i -th radiologist measurement of the k -th case representing each case, we used the following formula with a function of a scaling factor d on a percent scale:

$$M'_{ik} = \overline{M}_k + (1 + d/100)(M_{ik} - \overline{M}_k).$$

Here, M_{ik} is the observed reading of the i -th radiologist i for the k -th case, and \overline{M}_k is the median of measurements for the k -th case over all 13 radiologists. Specifically, we adjusted the inter-observer variability by assigning different values of the factor d to the deviation $M_{ik} - \overline{M}_k$ for each radiologist. We assigned $d = 40, 0$, and -40 so that the generated new M'_{ik} measurement data represent those with increased, observed, and decreased inter-observer variability, respectively; the deviation of each radiologist-level measurement from the case-specific mean was increased, unchanged, or decreased by $d\%$.

Case with Increased Inter-observer Variability													
Set	Radiologist												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	5.14	4.20	3.52	3.71	4.46	4.06	4.65	4.19	4.62	4.23	4.46	5.45	4.79
2	2.87	2.47	2.79	2.91	2.23	2.14	2.52	3.07	2.80	3.15	2.35	2.68	2.55
3	2.23	1.84	1.98	1.91	1.88	1.64	1.67	2.38	2.38	2.12	1.70	1.98	1.74
4	6.86	5.36	6.14	6.51	6.51	4.97	4.90	5.63	6.06	6.73	6.66	6.58	5.75
5	3.14	2.27	2.12	2.57	2.86	2.46	2.32	3.30	2.86	3.07	2.57	3.37	2.68
6	3.46	3.44	4.13	3.62	3.40	3.34	4.10	3.61	3.06	3.36	3.99	3.18	3.61
7	4.40	3.95	4.63	4.01	3.79	3.86	4.63	4.18	3.91	4.21	4.64	3.81	4.28
8	2.03	2.48	2.06	1.88	2.03	1.56	1.89	1.92	1.56	2.00	2.07	2.00	2.07
9	4.15	4.29	4.96	5.19	4.15	4.01	4.22	4.36	3.61	5.47	4.57	4.26	4.59
10	1.70	1.47	1.12	1.46	1.35	1.60	1.60	1.50	1.56	1.68	1.50	1.58	1.68

Case with Observed Inter-observer Variability													
Set	Radiologist												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	4.95	4.28	3.79	3.93	4.46	4.18	4.60	4.27	4.58	4.30	4.46	5.17	4.70
2	2.82	2.53	2.76	2.85	2.36	2.30	2.57	2.96	2.77	3.02	2.45	2.68	2.59
3	2.14	1.86	1.96	1.91	1.89	1.72	1.74	2.25	2.25	2.06	1.76	1.96	1.79
4	6.65	5.58	6.14	6.40	6.40	5.30	5.25	5.77	6.08	6.56	6.51	6.45	5.86
5	3.01	2.39	2.28	2.60	2.81	2.52	2.42	3.12	2.81	2.96	2.60	3.17	2.68
6	3.46	3.45	3.94	3.58	3.42	3.38	3.92	3.57	3.18	3.39	3.84	3.26	3.57
7	4.34	4.02	4.50	4.06	3.90	3.95	4.50	4.18	3.99	4.20	4.51	3.92	4.25
8	2.02	2.34	2.04	1.91	2.02	1.68	1.92	1.94	1.68	2.00	2.05	2.00	2.05
9	4.19	4.29	4.77	4.93	4.19	4.09	4.24	4.34	3.80	5.13	4.49	4.27	4.50
10	1.66	1.50	1.25	1.49	1.41	1.59	1.59	1.52	1.56	1.65	1.52	1.58	1.65

Case with Decreased Inter-observer Variability													
Set	Radiologist												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	4.75	4.35	4.05	4.14	4.46	4.29	4.54	4.34	4.53	4.36	4.46	4.88	4.60
2	2.76	2.59	2.72	2.78	2.48	2.45	2.61	2.84	2.73	2.88	2.54	2.68	2.62
3	2.04	1.88	1.94	1.91	1.89	1.79	1.80	2.11	2.11	2.00	1.82	1.94	1.83
4	6.45	5.81	6.14	6.30	6.30	5.64	5.61	5.92	6.11	6.40	6.37	6.33	5.98
5	2.88	2.51	2.44	2.63	2.76	2.58	2.52	2.94	2.76	2.85	2.63	2.97	2.68
6	3.46	3.45	3.74	3.53	3.43	3.41	3.73	3.52	3.29	3.41	3.68	3.34	3.52
7	4.27	4.08	4.37	4.11	4.01	4.04	4.37	4.18	4.06	4.19	4.38	4.02	4.22
8	2.02	2.21	2.03	1.95	2.02	1.81	1.96	1.97	1.81	2.00	2.03	2.00	2.03
9	4.23	4.29	4.58	4.68	4.23	4.17	4.26	4.32	4.00	4.80	4.41	4.28	4.42
10	1.62	1.52	1.37	1.51	1.47	1.57	1.57	1.53	1.56	1.61	1.53	1.57	1.61

Supplementary 3.**Average percent systematic difference using pairwise approach**

The pairwise average percent systematic difference δ_{ij} was calculated for Bland-Altman analysis. The measure is based on the average difference in measurement between any pair of the i-th and j-th radiologists for the k-th cases as follows.

$$\delta_{ij} = \frac{2}{K} \sum_{k=1}^K \frac{M_{ik} - M_{jk}}{M_{ik} + M_{jk}}$$

Here, K is the number of cases (in our study $K = 10$), and M_{ik} is a measurement value of the i-th radiologist for the k-th case.

Bland-Altman outlier scores with standard and fixed-limit

The standard Bland-Altman outlier scores Y_{2SD} is reliant on the percentage of pairwise measurement difference less than 2 standard deviations. Similarly, the Bland-Altman scores $Y_{20\%}$ with 20% fixed limit is reliant on the percentage of pairwise measurement difference less than 20% and calculated as follows:

$$Y_{2SD} = \frac{(N-2)!}{N!} \sum_{i=1}^N \sum_{j=1}^N 1 \left(\frac{|M_{ik} - M_{jk}|}{M_{ik} + M_{jk}} < 2SD \right)$$

$$Y_{20\%} = \frac{(N-2)!}{N!} \sum_{i=1}^N \sum_{j=1}^N 1 \left(\frac{|M_{ik} - M_{jk}|}{M_{ik} + M_{jk}} < 0.2 \right)$$

Here, $1(A)$ is an indicator function whose value is 1 if A is true and 0 otherwise. N represents the number of radiologists. The fixed-limit Bland-Altman outlier scores were based on the percentage of pairs where a pair of radiologists reviewed the same CT image set and resulted in measurements that differ by less than 20%.