BMC Medical Imaging

RESEARCH

Open Access

Feasibility of an artificial intelligence system for tumor response evaluation

Nie Xiuli[1†], Chen Hua[2†], Gao Peng[3], Yu Hairong[1], Sun Meili[2] and Yan Peng[2*]

Abstract

Purpose The objective of this study was to evaluate the feasibility of using Artificial Intelligence (AI) to measure the long-diameter of tumors for evaluating treatment response.

Methods Our study included 48 patients with lung-specific target lesions and conducted 277 measurements. The radiologists recorded the long-diameter in axial imaging plane of the target lesions for each measurement. Meanwhile, AI software was utilized to measure the long-diameter in both the axial imaging plane and in three dimensions (3D). Statistical analyses including the Bland-Altman plot, Spearman correlation analysis, and paired *t*-test to ascertain the accuracy and reliability of our findings.

Results The Bland-Altman plot showed that the AI measurements had a bias of -0.28 mm and had limits of agreement ranging from −13.78 to 13.22 mm ($P=0.497$), indicating agreement with the manual measurements. However, there was no agreement between the 3D measurements and the manual measurements, with $P<0.001$. The paired *t*-test revealed no statistically significant difference between the manual measurements and AI measurements ($P=0.497$), whereas a statistically significant difference was observed between the manual measurements and 3D measurements ($P<0.001$).

Conclusions The application of AI in measuring the long-diameter of tumors had significantly improved efficiency and reduced the incidence of subjective measurement errors. This advancement facilitated more convenient and accurate tumor response evaluation.

Keywords RECIST 1.1, Artificial intelligence, Tumor measurement

Introduction

The evaluation of tumor response is of significant importance in both clinical trials and standard cancer treatments. The Response Evaluation Criteria in Solid Tumors (RECIST) criteria has been widely adopted for evaluating response in solid tumors [1]. Recently, RECIST 1.1 has emerged as the gold standard for evaluating treatment response in solid tumors [2].

Nevertheless, it is worth noting that significant variations exist among individuals when utilizing the RECIST criteria for tumor response evaluation [3]. The potential instability of evaluation results can be attributed to differences in the training, supervision, and quality control

†Nie Xiuli and Chen Hua contributed equally to this work.

*Correspondence:
Yan Peng
yanpeng5325126@126.com
[1]Department of Radiology, Jinan Central Hospital, Shandong First Medical University, Jinan, China
[2]Department of Oncology, Jinan Central Hospital, Shandong First Medical University, Jinan, China
[3]Department of Radiology, Jiaozhou Hospital of Tongji University Dongfang Hospital, Tongji University, Qingdao, China

measures implemented by radiologists across different centers [4]. The evaluation can also be influenced by the subject's clinical information, potentially introducing bias. Moreover, factors such as off-duty periods, turnover, sick leaves, and other circumstances that necessitate the replacement of radiologists at the center might affect its consistency.

A Blinded Independent Review Committee (BIRC) has been proposed as an effective solution to the aforementioned issues. However, it is essential to acknowledge that using a BIRC might introduce bias due to informative censoring, which occurs when unconfirmed locally determined progressions are censored [5]. Additionally, the process of reading films in batches at the individual subject level during the trial can result in a longer duration. Randomized clinical trials have shown a substantial discrepancy rate between BIRC and local reviews [6–8]. Therefore, while a BIRC might be considered as an auditing tool, its universal implementation can not be recommended [5].

The development of a tumor evaluation method that is stable, uniform, and efficient while minimizing subjective bias is an urgent issue in clinical practice. Notably, the implementation of Artificial Intelligence (AI) products is anticipated to automate measurements and replace subjective evaluation, thereby reducing bias. There are a plethora of software tools currently available, however, a significant number of them lack clinical validation and have not undergone multi-institutional external validation.

This study seeks to evaluated the feasibility of utilizing artificial intelligence for 2D measurements and their comparability to manual 2D measurements, an area that remains underexplored in real-world clinical practice with AI tools. To this end, we utilize a commercially available advanced AI-driven tool designed to rapidly and accurately identify potentially suspicious lesions in lung CT images. The system employs sophisticated AI algorithms not only to identify these lesions but also to provide a qualitative analysis of lesion characteristics, including size, location, density, and nature. Furthermore, the selected AI system automatically calculates parameters defined by the RECIST 1.1 criteria, such as the long-diameter of tumors in both the axial plane and three-dimensional (3D) space. This feature enables an in-depth comparison between 3D tumor volume measurements and traditional two-dimensional (2D) measurements, which facilitated further investigation into the potential advantages and limitations of AI in clinical imaging.

Methods

Patient cohort

Between July 2018 and March 2023, we conducted a comprehensive screening of patients participating in clinical trials at Jinan Central Hospital, Shandong First Medical University. In order to be eligible for inclusion, patients had to meet specific criteria, including the availability of complete data and the presence of target lesions confined to the lungs. Exclusion criteria encompassed patients who declined enrollment or opted out of treatment after successful enrollment, as well as those unable to obtain tumor response evaluation results. This study received approval from both the Research Council and Ethics Committee. All patients provided written informed consent to use of their imaging data.

Image data acquisition

Contrast-enhanced CT examinations were conducted using a dual-source CT scanner (SOMATOM Force, Siemens, Germany). The layer thickness was set at 5 mm, with thin layer reconstruction at 1 mm. Specifically, we chose the images from the arterial phase (with a layer thickness of 1 mm) and uploaded them to the AI CT lung workflow (R12.9). Within approximately 5–10 min, the main interface displayed the attribute information of the lesions in the form of a focal list. As shown in Fig. 1, the primary interface featured a focal list that presented attribute information pertaining to the lesions. Upon selecting a specific lesion, additional information regarding the lesion, such as the long-diameter in axial imaging plane identified by artificial intelligence, the long-diameter of the tumor in 3D view, tumor volume, tumor weight, surface area, CT value, solid proportion, compactness, sphericity, entropy and other parameters.

Data collection

The long-diameter measured by AI and the long-diameter observed in the 3D view, along with the tumor volume measured by AI, were recorded. Additionally, the patient's age, gender, manually measured long-diameter, tumor response evaluation results for the target lesions, and overall tumor evaluation results were obtained from the tumor response evaluation record tables. Each interpretation of the diameter measurement was conducted by an independent radiologist possessing a minimum of 10 years of experience.

Statistical analysis

Bland-Altman plots were used to evaluate the agreement between various measurement methods. Spearman correlation analysis was conducted to assess the correlation among the different measurements methods. Paired *t*-tests were employed to identify discrepancies between the different measurement methods. Bland-Altman plots

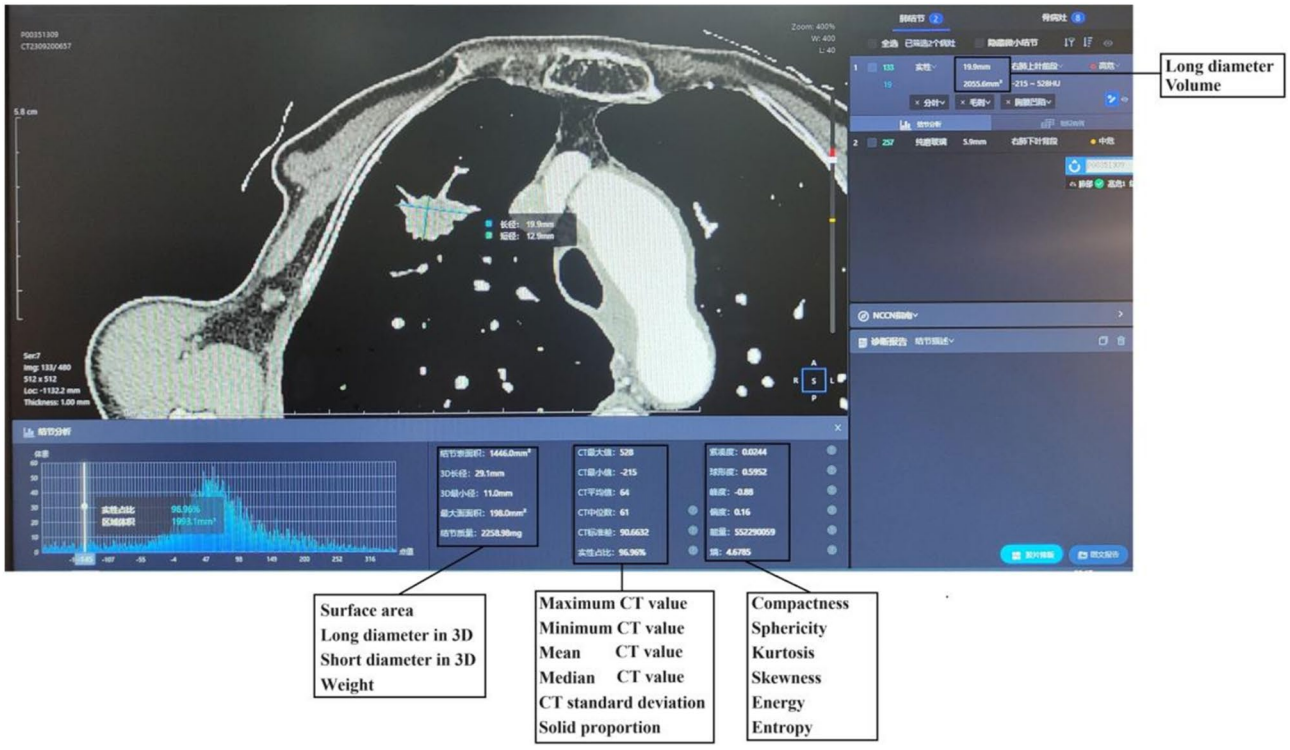

Fig. 1 The interface displayed of the AI CT lung workflow. This AI software can automatically measure of various parameters, including the long-diameter in the axial imaging plane, long-diameter in the three-dimensional view, tumor surface area, tumor volume, CT value, tumor weight and others. Since the software was originally in simplified Chinese, we provided an English translation for the figure

Table 1 Patient characteristics

Patients	Number *N* = 48	%
Age		
< 70	20	41.67%
≥ 70	28	58.33%
Gender		
Male	26	54.17%
Female	22	45.83%
Tumor types		
NSCLC	39	81.25%
SCLC	1	2.08%
Breast cancer	4	8.33%
Esophageal cancer	1	2.08%
Colorectal cancer	3	6.25%
Atelectasis		
With	11	22.92%
Without	37	77.08%
Immunotherapy		
Included	7	14.58%
Not included or unknown	41	85.42%

and correlation analyses were generated using MedCalc Software (version 11.4.2.0), while SPSS 27.0 software was used for paired *t*-tests, and scatter charts. GraphPad Prism software (v8.0) was used to generate heatmaps. A significance level of $P<0.05$ was considered statistically significant.

Results

Study population

A total of 48 patients who met the inclusion criteria were included in the study, as shown in Table 1. Among these patients, 20 were below the age of 70 while 28 were aged 70 years or older. The gender distribution consisted of 26 males and 22 females. Regarding the tumor category, 39 patients were diagnosed with non-small cell lung cancer (NSCLC), one patient with small cell lung cancer (SCLC), four patients with breast cancer, one patient with esophageal cancer, and three patients with colorectal cancer. Atelectasis was observed in 11 patients, while the remaining 37 patients did not exhibit this condition. Additionally, seven patients were identified as having received immune checkpoint inhibitors (ICIs).

Correlation and agreement of different measurement methods

According to the findings presented in Fig. 2A, the long-diameter measured by manual, AI and 3D showed a strong correlation with the tumor volume, as indicated by correlation coefficients of 0.845, 0.876, and 0.888, respectively, and all with $P<0.001$. These results suggested a

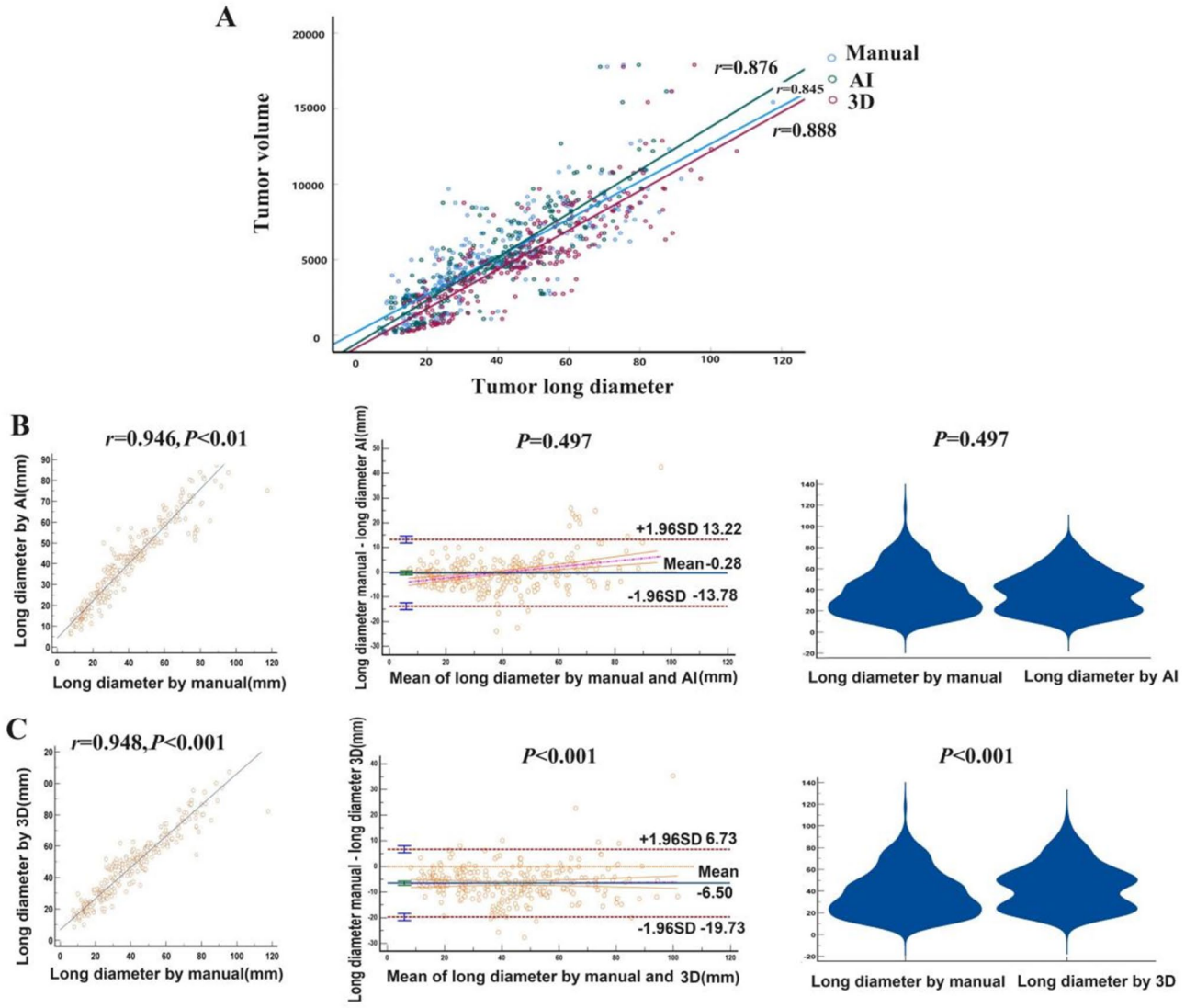

Fig. 2 Correlation and agreement of three measurement methods. **A**) Analysis of correlation between long-diameter measured by different methods and tumor volume. (**B**) Correlation and agreement between manual measurements and AI measurements. Spearman correlation analysis showed a close correlation between manual measurements and AI measurements. Bland-Altman plots represented the degree of agreement between manual and AI measurements. The paired *t*-test showed no statistical difference between the two measurement methods. (**C**) Correlation and agreement between manual measurements and 3D measurements. Although correlation analysis showed a strong correlation between manual and 3D measurements, Bland-Altman plots and the paired *t*-test revealed a statistical difference

close relationship between those three measurement methods and tumor volume. Additionally, the correlation analysis in Fig. 2B demonstrated a strong correlation between manual measurements and AI measurements ($r=0.946$, $P<0.001$). There was a significant agreement observed between manual measurements and AI measurements, as determined by Bland-Altman analysis ($P=0.497$). The analysis revealed a bias of -0.28 mm, with limits of agreement ranging from −13.78 to 13.22 mm. Furthermore, the paired *t*-test also confirmed no statistical difference between the two measurements methods ($P=0.497$). As shown in Fig. 2C, although manual measurements and 3D measurements exhibited a strong correlation ($r=0.948$, $P<0.001$), and there was no statistical agreement between manual measurements and 3D measurements, as determined by Bland-Altman analysis ($P<0.001$). The Bland-Altman plots revealed a bias of -6.50 mm, with limits of agreement ranging from −19.73 to 6.73 mm. In addition, there was a statistical difference between the two measurement methods ($P<0.001$).

The difference in long diameters of different measurement methods

The study presented in Fig. 3 depicted difference ratios between manual measurements, AI measurements, and 3D measurements. The corresponding formulation

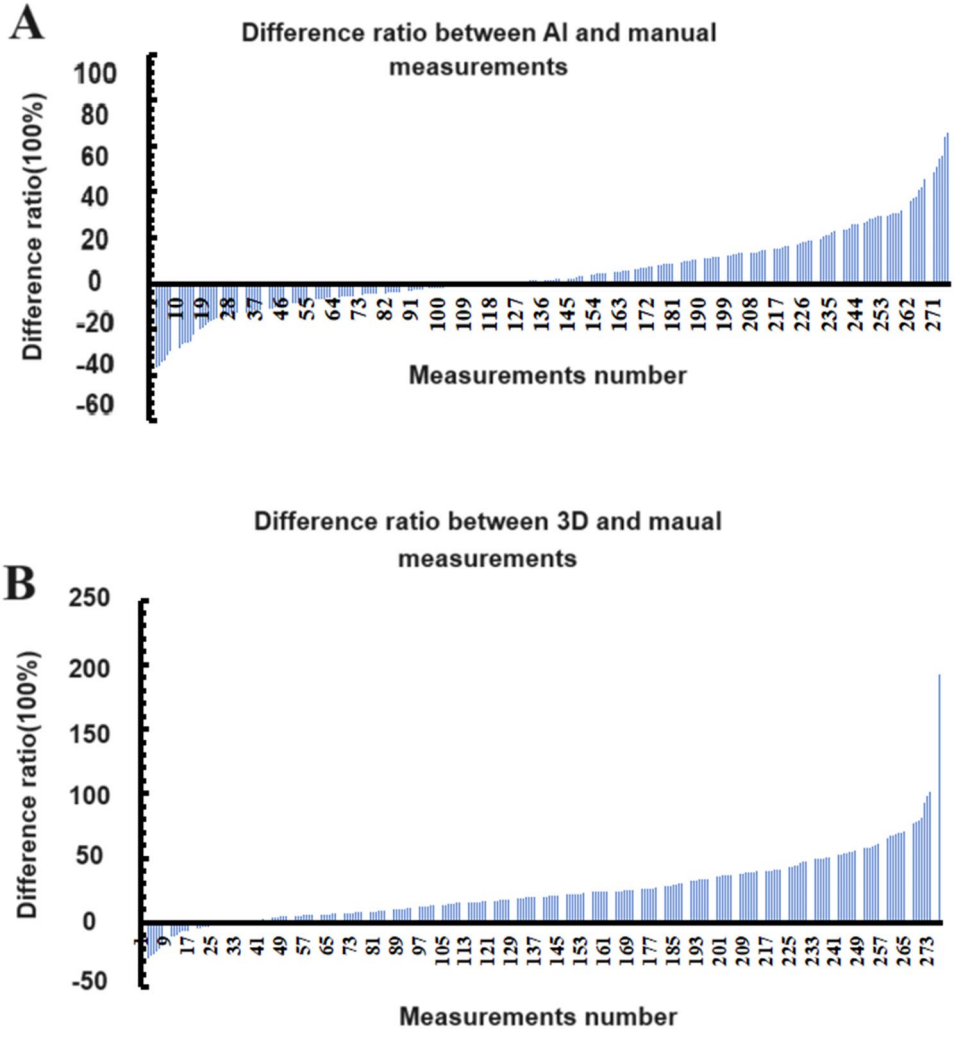

Fig. 3 The difference ratios in the long-diameter measurements between manual, AI and 3D methods, The waterfall plot depicted the difference ratios between the long-diameter measured manually and measured by AI (**A**) or 3D methods (**B**)

was $\frac{long\,diameter\,(AI)-long\,diameter\,(manual)}{long\,diameter\,(manual)}$ (difference ratios between manual measurements and AI measurements) or $\frac{long\,diameter\,(3D)-long\,diameter\,(manual)}{long\,diameter\,(manual)}$ (difference ratios between manual measurements and 3D measurements). In comparison to manual measurements, the median difference ratio of AI measurements was found to be 4.70% ± 18.30%. Out of the 277 manual measurements, 115 (41.52%) yielded larger long-diameters than AI measurements, while 5 (1.81%) resulted in the same long-diameter, manual measurements exhibited smaller long-diameters in 157 (56.68%) measurements. However, the utilization of the 3D measurements method generally resulted in greater long-diameters when compared to manual measurements. There were 244 (88.09%) measurements of larger long-diameters and only 33 (11.91%) measurements of smaller long-diameters, with a median value of 24.66% ± 27.00%.

The horizontal axis represented the number of measurements (ordered by difference ratios), and the vertical axis represented the difference ratios.

Tumor response evaluation results with different measurement methods

The objective of measuring the long-diameter was to evaluate the tumor response in accordance with the RECIST criteria, we examined whether variations were present in the evaluation of the tumor based on different measurement methods. The tumor response evaluation, as depicted in Fig. 4, demonstrated variations in measurements methods. When evaluated using the AI measurement method, 19 (6.86%) measurements exhibited inconsistencies compared to manual measurements, whereas 44 (15.88%) measurements displayed inconsistencies between the 3D measurement method and manual measurement method.

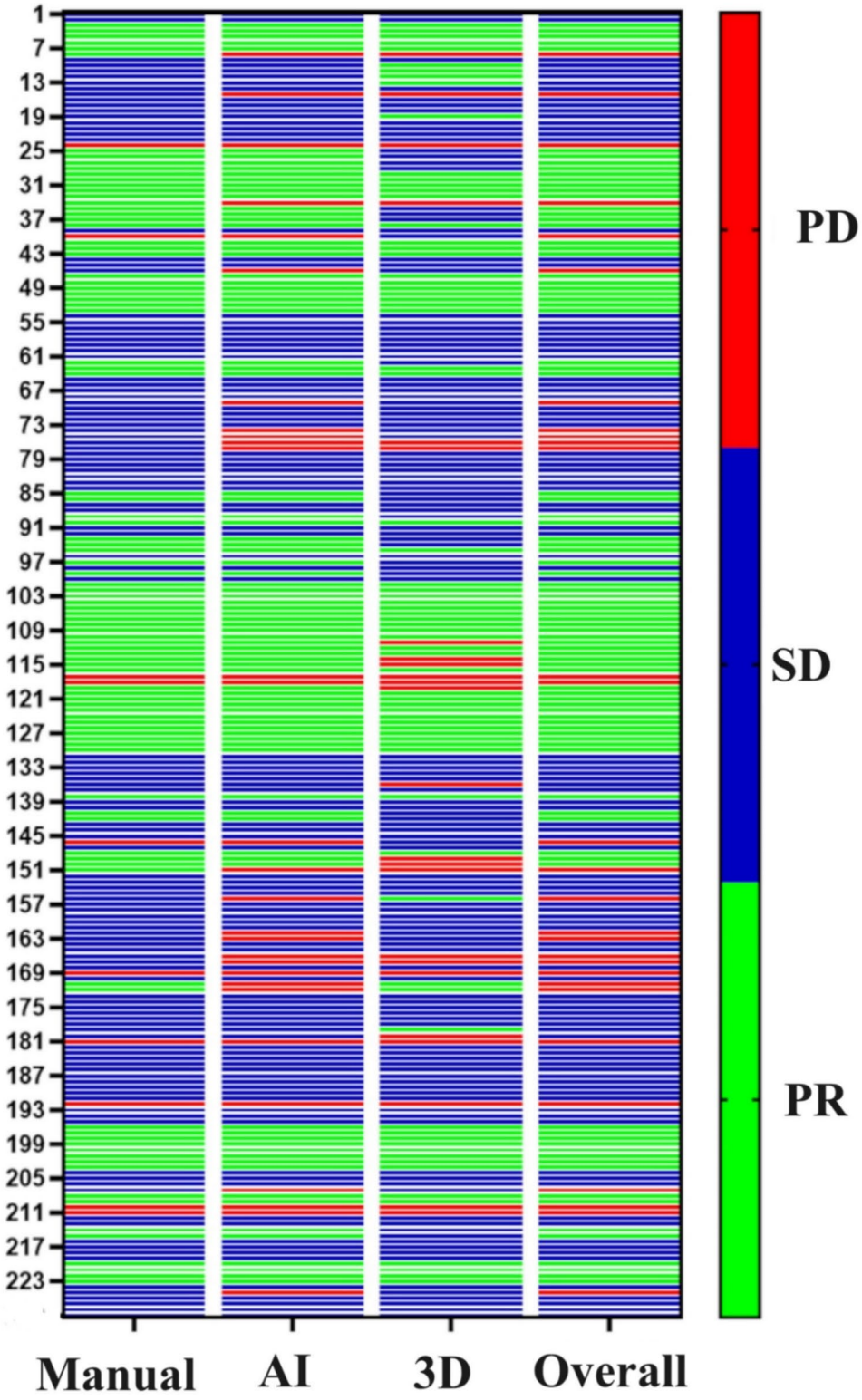

Fig. 4 Tumor response evaluation results by different measurement methods. The heatmap showed the tumor response evaluation by different measurements. The columns from left to right represented manual measurements of target lesions response evaluation, AI measurements of target lesions, 3D measurements of target lesions and overall tumor response. The evaluation of target lesions involved measuring their size, while the overall evaluation encompasses monitoring for the progression of non-target lesions or the emergence of new lesions. PD: progressive disease; SD: stable disease; PR: partial response

Correlation and agreement between different measurement methods for patients with atelectasis

Notably, the Bland-Altman plot revealed a discrepancy between manual measurements and AI measurements ($P=0.019$) in a patient with atelectasis. The plot reveals a bias of -1.09 mm, with limits of agreement ranging from −8.20 to 6.02 mm. Additionally, there was a significant disagreement between manual measurements and 3D measurements ($P<0.001$), with a bias of -7.04 mm and limits of agreement from −14.70 to 0.62 mm (Fig. 5A). The scatter plot distribution indicated that both AI measurements and 3D measurements exhibit greater long-diameters than manual measurements (Fig. 5B).

.

Discussion

Our research demonstrated that the utilization of this AI assistant detection system could effectively evaluate tumor response. This could reduce human error, improve standardization, and increase efficiency.

In recent years, there had been a growing utilization of AI systems to assist radiologists in their interpretations and mitigate reader inconsistencies. Additionally, several studies had explored the application of AI technology for tumor evaluation. For instance, one study employed Picture Archiving and Communication System (PACS) and Lesion Management Solutions System (LMS, Median Technologies, Valbonne Sophia Antipolis, France) software to identify intra- and inter-observer variability in measuring target lesions [9]. Furthermore, the Autocontour software from GE Healthcare was

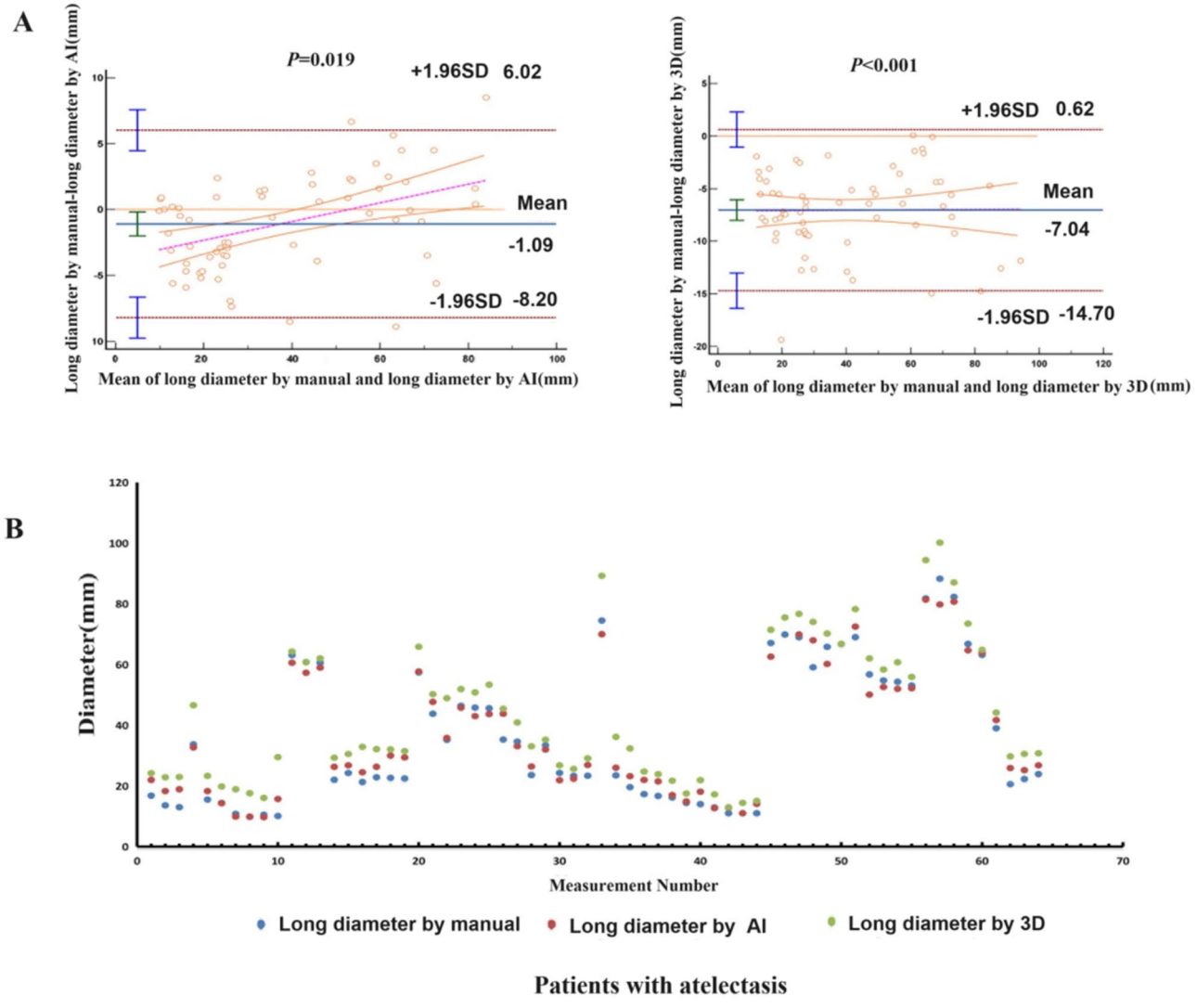

Fig. 5 Correlation and agreement of three measurement methods in patients with atelectasis. (**A**) The Bland-Altman plot demonstrated the disagreement between manual measurements and AI measurements, and 3D measurements. (**B**) The scatter plot indicated long-diameters of 3D and AI measurements were significantly larger than those measured manually

used for semi-automated volume analysis of malignant liver tumors [10]. Meanwhile, Myrian Intrasense artificial intelligence software (Paris, France) was trained to accurately measure mesothelioma tumor volume using CT images [11]. The CT lung assistant detection system implemented in our hospital was initially designed for the identification of benign and malignant tumors. This system possesses the capability to automatically measure the long-diameter of tumors, a feature developed by us for tumor response evaluation. Our investigation indicated that this system had the potential to evaluated tumor response stably.

The limitations of RECIST 1.1 attributed to its inherent design [12–15] and the presence of intra-observer and inter-observer measurements errors [16–19]. In an effort to evaluate the variability of lung tumor measurements, Oxnard et al. evaluated variability of lung tumor measurements using repeat CT scans performed within 15 min, they found median increase and decrease in tumor measurements were 4.3% and 4.2%, respectively [16]. The other study explored intra- and inter-observer variability of tumor responds. It was be found that 40% of major disagreements (PD or Non-PD) occurred and minor disagreements (disagreements between complete response (CR), PR, or SD) in 10.5% of the reviewed files [19]. They believed errors in tumor measurements was one reason for disagreements occurred. Reducing intra- and inter-observer measurements errors was an important method to maintain the stability of tumor evaluation results. This CT lung artificial intelligence system had the potential to address both intra-observer errors, which occur when a radiologist makes different measurements, and inter-observer errors, which arise from discrepancies between measurements taken by different radiologists.

This study determined that the long-diameter in the axial imaging plane measured by AI was a more effective alternative to manually measured than the long-diameter measured by 3D methods, although it was worth noting that the long diameter in 3D exhibited a strong correlation with tumor volume. As emphasized in the RECIST 1.1 update and clarification [1], it was advised to utilize the axial imaging plane in all instances of CT scans for the sake of uniformity and convenient measurements, particularly in situations where reconstructions or advanced workstations may not be universally accessible. It was acknowledged that alternative planes might accurately depict the tumor's long axis, but due to potential challenges in consistently and reliably measuring across different CT acquisition parameters over time, utilizing the axial plane was recommended.

There were some limitations in this study. Firstly, as we mentioned, it became apparent that this commercial AI detection system demonstrated its inadequacy for individuals diagnosed with atelectasis, as the AI measurements of the long-diameter were significantly greater than those obtained through manual measurement due to the inclusion of atelectasis components within the target lesion. This discrepancy had the potential to complicate the accurate evaluation of tumor response. In addition to atelectasis, AI measurements were constrained by the presence of fibrosis or necrotic tissue, particularly in cases where non-viable residual masses persisted following treatments such as radiotherapy and ablation, resulting in an underestimation of treatment response [20]. This commercial AI detection system was a traditional artificial intelligence, the alternative form of artificial intelligence was deep learning, which could automatically learn feature representations from data without the need for prior definition by human experts [21], which had the ability to recognize atelectasis and necrosis. Secondly, this study just evaluate the consistency between manual measurements and AI measurements, without evaluating the consistency of decision-making processes. Specifically, the tumor evaluation still required manual calculation of the total sum of diameters, which was then interpreted in accordance with the RECIST 1.1 criterion, not full automation. Thirdly, this AI software was exclusively designed as an artificial intelligence system for the interpretation of lung lesions, thereby lacking the capability to evaluate lesions in other body organs, which constituted a significant constraint.

In addition to measuring long-diameter of the tumors, this AI algorithm successfully acquired various other variables, including lesion surface area, lesion weight, CT value, compactness, sphericity, and firmness ratio, and others. Integrating these parameters with tumor diameters had the potential to enhance the precision of evaluating tumor response following treatment.

Conclusions

Our investigation demonstrated the potential of employing AI as a substitute for tumor response evaluation. This approach could enhance the efficiency of tumor evaluation while mitigating intra-observer and inter-observer errors, thereby warranting consideration for broader implementation in clinical practice.

Acknowledgements

Thanks to Dr. Gao Zhen, Zhang Chunling, Zhou Peng for their measurements work on the target lesions.

Author contributions

N.XL.: Writing - Original draft, investigation; C.H.: Screened patients according to the inclusion and exclusion criteria; G.P.: Artificial intelligence data collection, curation and visualization; Y.HR. : Artificial intelligence data collection, collation and visualization; S.ML.: Project administration; Y.P.: Supervision, review, formal analysis, editing and funding acquisition. All authors reviewed the manuscript.

Funding
Youth Project of Natural Science Foundation of Shandong Province (ZR2022QH187);

Data availability
Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate
The study adhered to the tenets of the Declaration of Helsinki and was approved by the Jinan Central Hospital, Shandong First Medical University Clinical Human Research Ethics Committee. All participants provided written informed consent to have their images.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 May 2024 / Accepted: 8 October 2024
Published online: 18 October 2024

References

1. Schwartz LH, Litière S, de Vries E, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1-Update and clarification: from the RECIST committee. Eur J Cancer. 2016;62:132–7.
2. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228–47.
3. Ko CC, Yeh LR, Kuo YT, Chen JH. Imaging biomarkers for evaluating tumor response: RECIST and beyond. Biomark Res. 2021;9(1):52.
4. Woo M, Heo M, Devane AM, Lowe SC, Gimbel RW. Retrospective comparison of approaches to evaluating inter-observer variability in CT tumour measurements in an academic health centre. BMJ Open. 2020;10(11):e040096.
5. Dodd LE, Korn EL, Freidlin B, Jaffe CC, Rubinstein LV, Dancey J, et al. Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense. J Clin Oncol. 2008;26(22):3791–6.
6. Geyer CE, Forster J, Lindquist D, Chan S, Romieu CG, Pienkowski T, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. N Engl J Med. 2006;355(26):2733–43.
7. Miller KD, Chap LI, Holmes FA, Cobleigh MA, Marcom PK, Fehrenbacher L, et al. Randomized phase III trial of capecitabine compared with bevacizumab plus capecitabine in patients with previously treated metastatic breast cancer. J Clin Oncol. 2005;23(4):792–9.
8. Thomas ES, Gomez HL, Li RK, Chung HC, Fein LE, Chan VF, et al. Ixabepilone plus capecitabine for metastatic breast cancer progressing after anthracycline and taxane treatment. J Clin Oncol. 2007;25(33):5210–7.
9. Muenzel D, Engels HP, Bruegel M, Kehl V, Rummeny EJ, Metz S. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. Radiol Oncol. 2012;46(1):8–18.
10. Dubus L, Gayet M, Zappa M, Abaleo L, De Cooman A, Orieux G, et al. Comparison of semi-automated and manual methods to measure the volume of liver tumours on MDCT images. Eur Radiol. 2011;21(5):996–1003.
11. Kidd AC, Anderson O, Cowell GW, Weir AJ, Voisey JP, Evison M, et al. Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: validation and comparison with modified RECIST response criteria. Thorax. 2022;77(12):1251–9.
12. Scher HI, Morris MJ, Kelly WK, Schwartz LH, Heller G. Prostate cancer clinical trial end points: RECISTing a step backwards. Clin Cancer Res. 2005;11(14):5223–32.
13. Chukwueke UN, Wen PY. Use of the Response Assessment in Neuro-Oncology (RANO) criteria in clinical trials and clinical practice. CNS Oncol. 2019;8(1):CNS28.
14. Costelloe CM, Chuang HH, Madewell JE, Ueno NT. Cancer Response Criteria and Bone metastases: RECIST 1.1, MDA and PERCIST. J Cancer. 2010;1:80–92.
15. Byrne MJ, Nowak AK. Modified RECIST criteria for assessment of response in malignant pleural mesothelioma. Ann Oncol. 2004;15(2):257–60.
16. Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, et al. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. J Clin Oncol. 2011;29(23):3114–9.
17. Stewart DJ, Kurzrock R. Fool's gold, lost treasures, and the randomized clinical trial. BMC Cancer. 2013;13:193.
18. Choi JH, Ahn MJ, Rhim HC, Kim JW, Lee GH, Lee YY, et al. Comparison of WHO and RECIST criteria for response in metastatic colorectal carcinoma. Cancer Res Treat. 2005;37(5):290–3.
19. Thiesse P, Ollivier L, Di Stefano-Louineau D, Négrier S, Savary J, Pignard K, et al. Response rate accuracy in oncology trials: reasons for interobserver variability. Groupe Français d'Immunothérapie of the Fédération Nationale Des Centres de Lutte Contre Le Cancer. J Clin Oncol. 1997;15(12):3507–14.
20. Grimaldi S, Terroir M, Caramella C. Advances in oncological treatment: limitations of RECIST 1.1 criteria. Q J Nucl Med Mol Imaging. 2018;62(2):129–39.
21. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017;550(7676):354–9.

Publisher's note