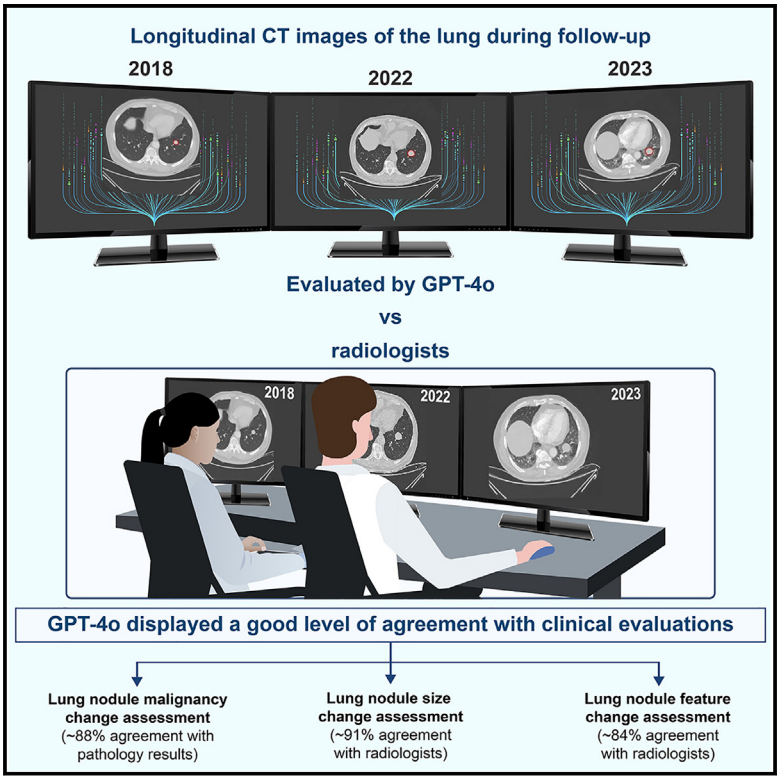
Cell Reports
Medicine

Article

Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images

Graphical abstract



Highlights

- GPT-4o evaluates malignancy progression in lung nodules in longitudinal CT images
- GPT-4o monitors dynamic changes in both nodule size and textural characteristics
- Comparing to deep learning models, GPT-4o offers interpretable predictions
- GPT-4o delivers precise radiological insights to support clinical decision-making

Authors

Yuqiang Mao, Nan Xu, Yanan Wu, ..., Dongmei Pei, Lina Zhang, Jiangdian Song

Correspondence

peidm1111@hotmail.com (D.P.), lnzhang@cmu.edu.cn (L.Z.), song.jd0910@gmail.com (J.S.)

In brief

In this study, Mao et al. utilize GPT-4o to analyze longitudinal CT follow-up images, enabling dynamic tracking of lung nodule malignancy probabilities and changes in size and characteristics. This approach provides valuable radiological insights with the potential to enhance clinical disease management.

Mao et al., 2025, Cell Reports Medicine 6, 101988
March 18, 2025
https://doi.org/10.1016/j.xcrm.2025.101988

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images

Yuqiang Mao,[1,7] Nan Xu,[2,7] Yanan Wu,[2,7] Lu Wang,[2,3,7] Hongtao Wang,[4,7] Qianqian He,[2] Tianqi Zhao,[5] Shuangchun Ma,[5] Meihong Zhou,[5] Hongjie Jin,[1] Dongmei Pei,[6,*] Lina Zhang,[5,*] and Jiangdian Song[2,8,*]

[1]Department of Thoracic Surgery, Shengjing Hospital of China Medical University, Shenyang, Liaoning 110004, China
[2]School of Health Management, China Medical University, Shenyang, Liaoning 110122, China
[3]Shengjing Hospital of China Medical University, Shenyang, Liaoning 110004, China
[4]Department of Hematology, Shengjing Hospital of China Medical University, Shenyang, Liaoning 110004, China
[5]Department of Radiology, The Fourth Affiliated Hospital of China Medical University, Shenyang, Liaoning 110032, China
[6]Department of Health Management, Shengjing Hospital of China Medical University, Shenyang, Liaoning 110004, China
[7]These authors contributed equally
[8]Lead contact
*Correspondence: peidm1111@hotmail.com (D.P.), lnzhang@cmu.edu.cn (L.Z.), song.jd0910@gmail.com (J.S.)
https://doi.org/10.1016/j.xcrm.2025.101988

SUMMARY

Large language models have shown efficacy across multiple medical tasks. However, their value in the assessment of longitudinal follow-up computed tomography (CT) images of patients with lung nodules is unclear. In this study, we evaluate the ability of the latest generative pre-trained transformer (GPT)-4o model to assess changes in malignancy probability, size, and features of lung nodules on longitudinal CT scans from 647 patients (547 from two local centers and 100 from a public dataset). GPT-4o achieves an average accuracy of 0.88 in predicting lung nodule malignancy compared to pathological results and an average intraclass correlation coefficient of 0.91 in measuring nodule size compared with manual measurements by radiologists. Six radiologists' evaluations demonstrate GPT-4o's ability to capture changes in nodule features with a median Likert score of 4.17 (out of 5.00). In summary, GPT-4o could capture dynamic changes in lung nodules across longitudinal follow-up CT images, thus providing high-quality radiological evidence to assist in clinical management.

INTRODUCTION

Large language models (LLMs), such as Chat generative pre-trained transformer (GPT) and Gemini, have stimulated the surge of natural language processing techniques that allow text comprehension and human-like response generation.[1] Prior LLMs have shown the potential to interpret radiological images, thus aiding in the detection of anomalies and making diagnostic decisions.[2–4] GPT-4o, which is the latest visual LLM released in May 2024 by OpenAI, expands these capabilities to encompass the processing of images, audio, and video, highlighting its potential use in multimodal medical applications. However, the use of video processing based on visual LLMs in clinical settings remains unexplored.

In the context of computed tomography (CT) image-based lung cancer analyses, advancements in deep learning (DL) have significantly impacted tumor screening and diagnosis.[5,6] A convolutional neural network (CNN)-based DL algorithm was developed to assess malignancy risk from low-dose CT scans, yielding an area under the curve (AUC) of 0.93.[7] Ardila et al. constructed a three-dimensional CNN model to analyze consecutive CT images for cancer risk prediction, achieving an AUC of 0.94.[8] In addition, a previous study reported no statistical significance between radiomics and a CNN model for lung nodule diagnosis; however, an increase in diagnostic accuracy was observed when the CNN model was integrated with clinical features.[9] The reader study of the aforementioned research using radiologists' assessments showed that, compared with clinical experts, the two DL algorithms[7,8] achieved better results when analyzing the CT examinations. However, the output from DL models often lacks interpretable reasoning, presenting challenges for clinical adoption despite proven performance metrics.[10] This limitation stems from the complex nature of DL algorithms,[11] where the internal decision-making process remains largely obscure to end users. Although traditional clinical reasoning follows clear diagnostic pathways that physicians can explain to patients, DL models typically provide predictions without revealing the specific features or patterns that led to their conclusions.

Recently, GPT models have shown the potential to automatically label oncological phenotypes and determine TNM stages based on radiology reports.[12–14] However, the ability to dynamically measure changes in the lung nodule size and characteristics across serial CT images is paramount to

assessing the progression of the malignancy risk and guiding timely clinical interventions, particularly for early-stage lung cancer.[15,16] Substantial clinical evidence has suggested a correlation between an increasing nodule diameter and the likelihood of malignancy.[17,18] Therefore, tracking dynamic nodule changes via follow-up CT evaluations provides insights that are valuable for the estimation of the potential malignancy risk.[19] Currently, this process necessitates a manual review of each subsequent CT image by radiologists; therefore, the assessment of nodular changes is time-consuming and susceptible to interobserver variability. This variability is particularly pronounced when subtle but crucial details, such as small bronchi involvement or vascular penetration—which are key indicators of malignancy progression—are present on an interim CT scan.[20,21] Automation of the assessment of dynamic changes in nodular characteristics could significantly enhance radiological efficiency and facilitate the early detection and timely management of patients at high risk for lung cancer.

This study aimed to leverage the advanced capabilities of GPT-4o to estimate the probability of lung nodule malignancy and dynamic changes in the nodule size and characteristics using longitudinal CT follow-up images of individual patients. By simulating the visual inspection procedures used by radiologists when reviewing videos of CT images, this study sought to evaluate the capacity of GPT-4o to accurately assess nodule characteristics over time and determine the potential of GPT-4o to enhance both diagnostic accuracy and efficiency in the context of lung cancer screening and clinical monitoring.

RESULTS

Patients

This multi-center retrospective study aimed to evaluate the performance of GPT-4o in detecting changes in malignancy probability, size, and features of patients with lung nodules on longitudinal CT follow-up images. The study flowchart with an example participant is shown in Figure 1. A total of 647 patients were retrospectively enrolled. Of these patients, 278 (140 benign cases; 138 malignant cases; mean age, 57 years; SD, 11.4 years; 89 men [32.0%]) were from the C1 dataset (Shengjing Hospital of China Medical University), 191 (92 benign cases; 99 malignant cases; mean age, 60 years; SD, 10.4 years; 72 men [37.7%]) were from the C2 dataset (the Fourth Affiliated Hospital of China Medical University), 100 (76 benign cases and 24 malignant cases; mean age, 62 years; SD, 4.8 years; 64 men [64.0%]) were from the National Lung Screening Trial (NLST) dataset, and 78 (mean age, 50 years; SD, 13.0 years; 51 men [65.4%]) were from the local lung cancer screening (LLCS) dataset (Table 1; Figure 2).

The patients enrolled in this study underwent a mean of 2.8 (SD, 1.2; range 2–10) CT examinations and an average follow-up interval of 286.2 days (SD, 491.2 days). The median tumor sizes of patients from the C1, C2, LLCS, and NLST datasets were 9.0 mm (interquartile range [IQR], 7.0–13.0), 9.5 mm (IQR, 7.0–14.0), 5.2 mm (IQR, 3.6–9.3), and 5.0 mm (IQR, 4.0–7.5), respectively, according to the initial CT images.

Evaluation of GPT-4o's diagnosis regarding the estimation of nodule malignancy

For the C1 and C2 datasets, the AUCs for the lung nodule malignancy estimations by GPT-4o based on only the first CT scan images were 0.75 (95% confidence interval [CI], 0.71–0.80; 278/278 patients) and 0.69 (95% CI: 0.64–0.74; 191/191 patients), respectively. When data from the first two follow-up evaluations were used, the AUCs improved to 0.86 (95% CI, 0.82–0.90; 278/278 patients, $p < 0.001$, DeLong test) and 0.88 (95% CI, 0.83–0.92; 191/191 patients, $p < 0.001$, DeLong test), respectively. The AUCs increased to 0.87 (186/278 patients, $p < 0.001$, DeLong test compared to the baseline) and 0.93 (59/191 patients, $p < 0.001$, DeLong test compared to the baseline), respectively, when the third CT images were integrated (Figures S1A and S1B). For the NLST dataset, the AUC of the estimation of nodule malignancy determined by GPT-4o was 0.72 (95% CI, 0.61–0.83) based on the initial CT scan image; however, it increased to 0.88 (95% CI, 0.80–0.96) and 0.92 (95% CI, 0.87–0.98) when the second and final CT images, respectively, were integrated (Figure S1C). For the LLCS dataset, the intraclass correlation coefficient (ICC) for the estimation of the probability of lung nodule malignancy performed by GPT-4o and that performed by the radiologists (based on the Lung-RADS [Lung Imaging Reporting and Data System] criteria) was 0.66 based on the initial CT images; however, it increased to 0.74, 0.80, and 0.90 when the second, third, and fourth CT images, respectively, were integrated (Figure S1D). There was no significant difference in the accuracy of GPT-4o in estimating the malignancy of lung nodules on CT images among participants in different gender subgroups ($p = 0.75$).

The evaluation scores for feature detection by GPT-4o determined by the six radiologists were averaged, resulting in median agreement scores of 3.51 (IQR, 3.05–4.10; mean score, 3.52; SD, 0.75) and 4.33 (IQR, 4.06–4.67; mean score, 4.34; SD, 0.48) for the C1 and C2 datasets, respectively (Figure 3). An average ICC of 0.53 for evaluating the inter-rater agreement was achieved when comparing the scores of the six radiologists. Additionally, the characterization of the nodule features by GPT-4o demonstrated strong agreement with the predefined radiological features annotated in the NLST dataset, resulting in an average accuracy of 0.84 (Table S1). Furthermore, we evaluated Molmo-7B's (a state-of-the-art open-source, multimodal vision-language model) performance in the NLST dataset.[22] Molmo-7B achieved an average accuracy of 0.67 for feature detection (Tables S2 and S3), which is lower than the accuracy observed with GPT-4o. In addition, the performance of comparative experiments using GPT-4o and Claude is presented in Table S4.

Evaluation of changes in the nodule size by GPT-4o

Across the C1, C2, LLCS, and NLST datasets, the nodule size measurements by GPT-4o achieved ICCs of 0.86 (95% CI, 0.83–0.88), 0.95 (95% CI, 0.94–0.96), 0.88 (95% CI, 0.85–0.91), and 0.93 (95% CI, 0.89–0.95), respectively, when compared to the manual measurements performed by the radiologists ($p < 0.001$). Pearson's correlation coefficients between the nodule size measurements performed by GPT-4o and radiologists were 0.86, 0.96, 0.88, and 0.93 for the C1, C2, LLCS, and NLST datasets, respectively, indicating significant agreement

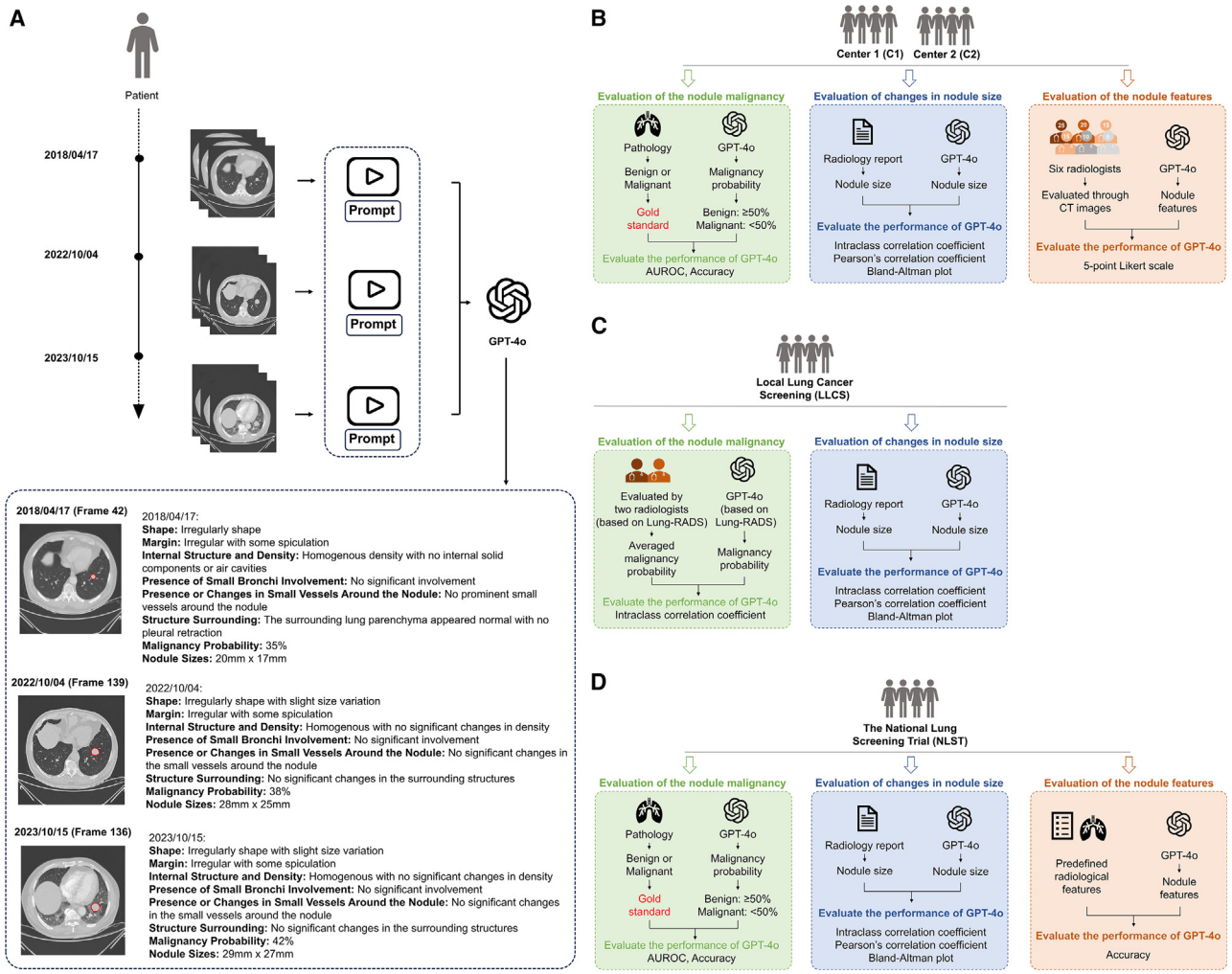


Figure 1. Study flowchart

(A) The overview of GPT-4o analysis used in this study. The follow-up computed tomography (CT) images of each patient were obtained from the picture archive and communication system and then transformed into a video with 20 frame images per second. These videos were input in GPT-4o to allow estimation of the probability of lung nodule malignancy, nodule size measurement, and nodule feature description. The results were then evaluated by six radiologists.
(B–D) The detailed evaluation process was illustrated in the C1 and C2 dataset (B), LLCS dataset (C), and NLST dataset (D). LLCS, local lung cancer screening; NLST, National Lung Screening Trial.

($p < 0.001$, Figure S2A). The Bland-Altman plots to illustrate the differences are presented in Figure S2B.

Additionally, the radiologists observed 105, 54, 51, and 44 patients from the C1, C2, LLCS, and NLST datasets, respectively, with increased nodule sizes (slope from 0.03 to 32.0) on the follow-up CT images. In contrast, 48, 36, 11, and 28 patients from the C1, C2, LLCS, and NLST datasets, respectively, had reduced nodule sizes (slope from −0.03 to −11.0). Compared to the accuracy of the radiologists' observations, GPT-4o achieved accuracy rates of 84.8% (89/105), 92.6% (50/54), 88.2% (45/51), and 86.4% (38/44) for tumor growth in the C1, C2, LLCS, and NLST datasets, respectively, as well as accuracy rates of 85.4% (41/48), 91.7% (33/36), 81.8% (9/11), and 85.7% (24/28) for tumor reduction in the C1, C2, LLCS, and NLST datasets, respectively. GPT-4o showed statistically significant better performance for malignancy estimation compared to using nodule size on CT images (Table S5). In addition, performance for nodule size evaluation and malignancy estimation using GPT-4o with different inputs is shown in Table S6.

Test-retest of the reproducibility of the evaluations

The test-retest experiments demonstrated an average reproducibility rate of 84.1% (an average of 71.5%, 78.6%, 83.0%, 85.2%, 92.5%, and 94.0% for the six radiologists) for the radiological nodule feature detection of GPT-4o based on the reviews of the two rounded assessments of 50 cases by the six radiologists. Additionally, the average reproducibility rates of the quantitative malignancy probability estimates were 82.0% (41/50) with the first CT images, 90.0% (45/50) with the second CT images, and 96.0% (48/50) with the third CT images.

To demonstrate the stability of GPT-4o in delineating nodule boundaries, a second round of experiment (Nov 11, 2024) for

Table 1. Demographic and clinical characteristics of enrolled participants

	C1		C2		LLCS	NLST	
Characteristics	Malignant	Benign	Malignant	Benign	All	Malignant	Benign
Number (%)	138 (49.6)	140 (50.4)	99 (52.0)	92 (48.0)	78 (100.0)	24 (24.0)	76 (76.0)
Age, yr, mean (SD)	59 (10.8)	56 (11.8)	60 (11.3)	59 (9.2)	50 (13.0)	63 (5.1)	62 (4.6)
Gender, male (%)	48 (34.8)	41 (29.3)	34 (34.3)	38 (41.3)	51 (65.4)	17 (70.8)	47 (61.8)
Size (median, in mm)	12 × 10	9 × 8	14 × 11	8 × 7	7 × 5	8 × 6	5 × 4
Slices, median (IQR)	357 (333, 381)	355 (327, 384)	340 (310, 385)	320 (280, 367)	366 (335, 387)	156 (130, 168)	152 (136, 163)
Video, time, mean (SD)	28.7 (8.0)	28.9 (6.5)	30.2 (11.2)	26.1 (12.9)	27.3 (8.8)	12.5 (1.8)	12.4 (1.5)

LLCS, local lung cancer screening; NLST, National Lung Screening Trial.

nodule boundary delineation was conducted 1 week after the initial round of experiment (Nov 4, 2024), under identical experimental conditions. Figure S3 presents the examples of GPT-4o's output from the two rounds of experiment for six patients, demonstrating the stability of the GPT-4o model. In addition, Video S1, Video S2, and Video S3 show the detailed training and testing process.

During the *ad hoc* experiment, six radiologists rated their willingness to use the model, reliance on the information provided, perceived potential for harm, missing content, and inappropriate content associated with the three models: GPT-4o, one local DL model, and one online DL model.[7] The average scores for GPT-4o (which provided both predictions and evidence) regarding the willingness to use (86.9% vs. 54.2% vs. 72.4%) and reliance (70.2% vs. 45.7% vs. 62.9%) were significantly higher than those of the two DL models (which provided only predictions of the probability). In addition, the average score for GPT-4o regarding the perceived potential for harm (19.2% vs. 53.9% and 54.6%) and missing content (35.6% vs. 79.3% and 64.2%) was lower than that of the two DL models (Table S7). While DL models typically output only a malignancy probability score, GPT-4o generates interpretable descriptions related to nodule characteristics and size changes. Since these descriptions may contain content that radiologists consider inappropriate, radiologists perceived the potential for GPT-4o to produce more inappropriate content than DL models' single probability output (20.7% vs. 10.0% and 10.0%).

DISCUSSION

This study demonstrates the ability of GPT-4o to evaluate the changes in lung nodule malignancy probability, nodule size, and feature on longitudinal CT follow-up images obtained from multiple centers and deliver high-quality radiological diagnostic evidence.

Compared to the pathological results from the C1, C2, and NLST datasets, the ability of GPT-4o to assess the probability of lung nodule malignancy significantly improved when follow-up CT scan images were incorporated compared to that obtained when only the initial CT scan was used. However, as shown in Figure S1A, the AUC of the C1 dataset decreased when the fourth CT scan was included, which may be attributed to the reduced sample size. By the fourth CT scan, the sample size in the C1 dataset for participants with benign and malignant cases had decreased from 278 to 56. A smaller sample size may lead to insufficient statistical power and decreased data representativeness, making it difficult for the model to precisely capture adequate data variability and distinguish between benign and malignant nodules.[23,24]

The study revealed that, compared to the radiologists' manual measurements, GPT-4o was able to detect lung nodule sizes across internal and external NLST datasets, with accuracy rates of 85.0%, 92.2%, 87.1%, and 86.1% for detecting nodule growth/reduction on images of sequential transverse CT sections on the four datasets in this study. Additionally, based on the radiologists' evaluations of the key nodule features,

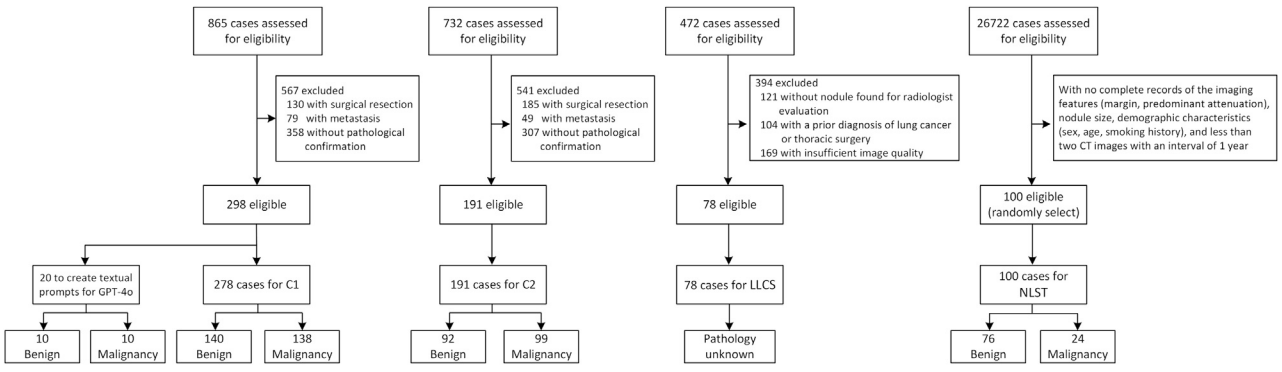


Figure 2. Patient enrollment flowchart
LLCS, local lung cancer screening; NLST, National Lung Screening Trial.

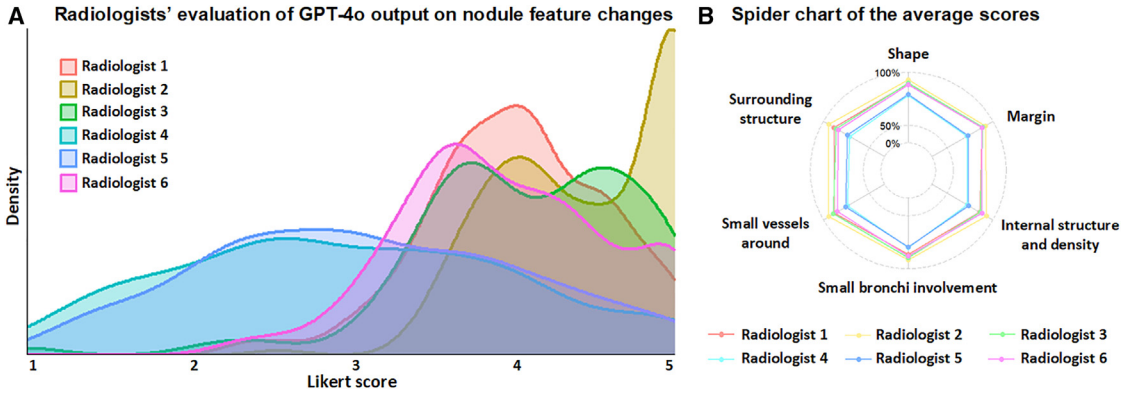


Figure 3. Radiologist's evaluation scores for feature characterization by the GPT-4o
(A) The density plot of six radiologists' Likert scores and (B) the spider chart of the average scores.

GPT-4o achieved a median Likert score of 4.17 (out of 5.00), thus demonstrating its ability to effectively identify signs of lung nodules on CT images. Moreover, for lung cancer screening setting, GPT-4o achieved good ICC for estimating the probability of lung nodule malignancy when compared to the evaluation of two radiologists, further supporting the potential of this approach to detect early lung cancer and improve the clinical management and monitoring of patients at risk for lung cancer.

Previous studies of DL methods have shown their capability to identify the probability of malignancy and malignancy-related nodular features using CT images.[8,25] Although a single CT scan image provides information regarding the nodule size, shape, and texture, longitudinal CT images reveal changes such as size increases, solid component growth, and the appearance of cystic components, indicating an increased risk of malignancy.[26–28] Conversely, size reductions and lesion absorption suggest a reduced risk of malignancy.[29–31] However, currently, these assessments rely on manual evaluations of follow-up CT images by radiologists, which are time-consuming and prone to interobserver variability. Our study demonstrated that the GPT-4o vision-language model can streamline this process by accurately identifying and directly monitoring lesion changes using videos of longitudinal CT images. Notably, our approach bypasses the need for complex, task-specific network architectures, such as U-net segmentation followed by classification networks.[32,33] Such approaches require specialized knowledge, thus limiting their practical implementation by radiologists. In contrast, GPT-4o eliminates the need for complex network engineering and offers a "plug-and-play" solution for monitoring lung nodules on longitudinal CT images. Furthermore, the findings of our *ad hoc* experiment suggest that this approach may enhance radiologists' willingness to utilize artificial intelligence techniques in clinical practice. Specifically, compared to the previously reported DL model,[7] GPT-4o was correlated with a 14.5% increase in willingness to use, 7.3% improvement in reliance on the information provided, 35.4% decrease in the perceived potential for harm, and 28.6% reduction in missing content. These findings highlight the potential of GPT-4o as a valuable tool that could assist radiologists in clinical practice.

Radiological nodule features are crucial for assessing the malignancy risk.[34–36] GPT-4o demonstrated the ability to identify subtle changes in nodule morphology over time. For instance, in Figure 4G, it accurately described a nodule's transition from "appears relatively round" on the initial CT scan to "shape has become slightly more irregular, with some mild protrusions, particularly in the third examination" on subsequent scans. This interpretation received a high credibility score of 5.0 out of 5.0 from the six expert radiologists. Additionally, GPT-4o detected evolving internal nodule characteristics, as illustrated in Figures 4D–7F. It noted a shift from "maintains uniform density" in the first two CT images to "increased internal density" in later scans. Furthermore, high agreement was observed in the assessment of "presence of small bronchi involvement" (Figure 4D), a potential indicator of malignancy.[37] GPT-4o achieved the same observations by radiologists (Figures 4G–4I), progressed from "no small bronchi involvement" on the initial image to "possible involvement of small bronchi adjacent to the nodule" on the interim scan and finally to "more pronounced involvement" on the last image. This aligns with the radiologists' evaluations over a 4-year follow-up period (score, 5.0/5.0) and the ultimate pathological confirmation of malignancy. These results suggest that GPT-4o could produce reports using language commonly used in the field of radiology.

Conclusion

In conclusion, this study demonstrated the capability of GPT-4o to emulate the ability of radiologists to monitor lung nodule characteristics and sizes and estimate the probability of malignancy using longitudinal CT follow-up images. Compared to pathological testing and manual evaluations by radiologists, GPT-4o achieved convincing accuracy when estimating the probability of malignancy and provided corresponding high-quality radiological evidence. These findings suggest the potential of GPT-4o to enhance the clinical management of patients at risk for lung cancer.

Limitations of the study

This study had some limitations. First, each CT scan was analyzed as a separate video input in GPT-4o, potentially reducing the efficiency and increasing the risk of technical issues, such as errors associated with video decoding and region of interest detection by Python, associated with the processing procedure of

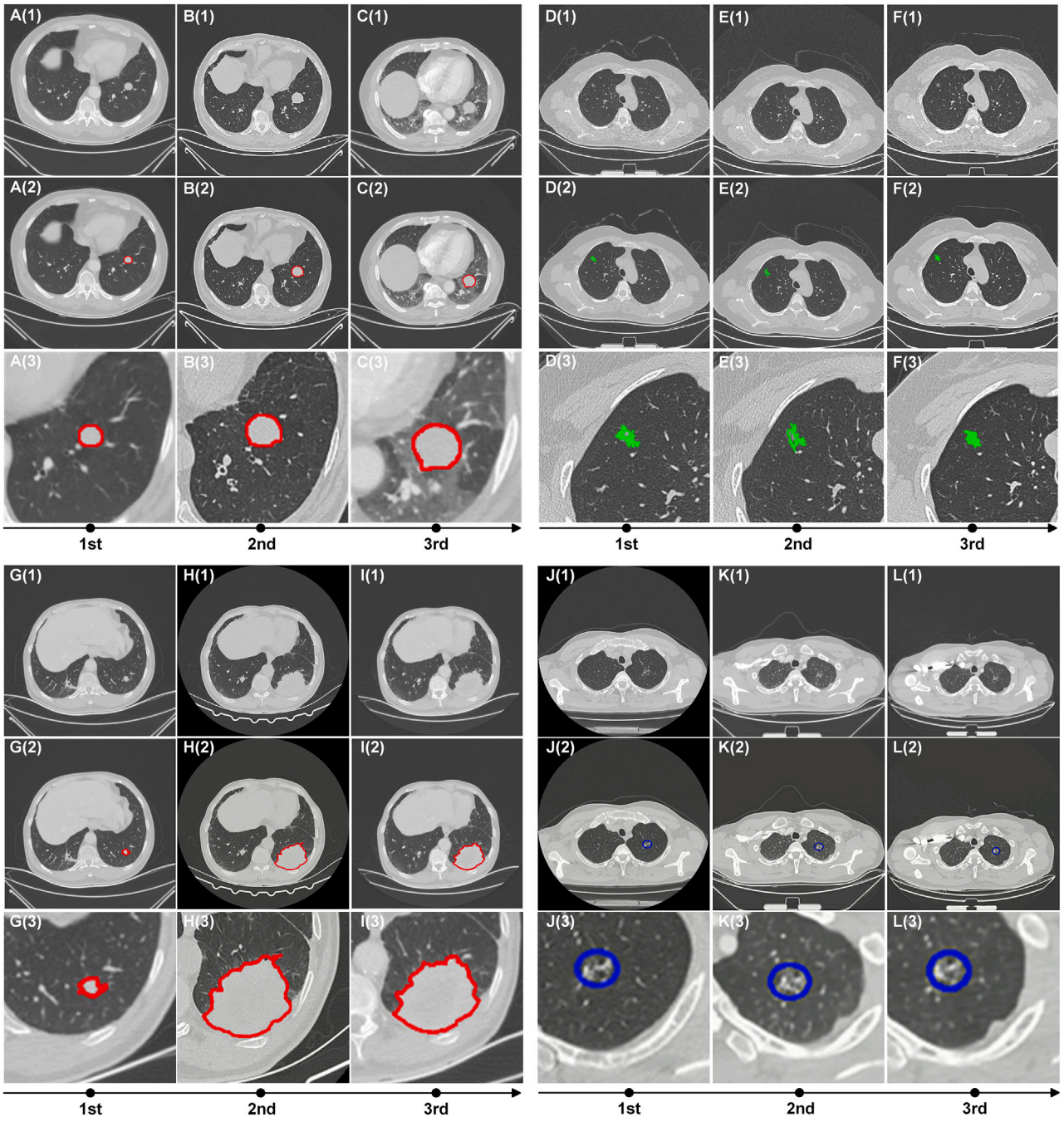


Figure 4. Examples of lung nodule margins marked by GPT-4o on follow-up CT images of four patients

(1) original CT images, (2) lung nodule margins marked by GPT-4o, and (3) magnified view of the marked regions highlighted in (2).

(A–C) Images of a 68-year-old man with a history of smoking and benign solid nodule in the basal segment of the left lower lobe. The measurements are as follows: initial image (April 17, 2018), 20 × 17 mm; follow-up image (October 4, 2022), 29 × 26 mm; and final image (October 15, 2023), 30 × 26 mm.

(D–F) Images of a 36-year-old woman with no history of smoking and a benign nodule in the right upper lobe. The measurements are as follows: initial image (November 18, 2021), 8 × 8 mm; follow-up image (January 24, 2022), 9 × 8 mm; and final image (May 24, 2023), 9 × 7 mm.

(G–I) Images of a 72-year-old man with a history of smoking and malignant nodule in the left lower lobe. The measurements are as follows: initial scan (February 28, 2018), 13 × 12 mm; follow-up image (June 22, 2021), 75 × 58 mm; and final image (June 28, 2022), 75 × 58 mm.

(J–L) Images of a 48-year-old man with a history of smoking and malignant nodule in the apicoposterior segment of the left upper lobe. The measurements are as follows: initial image (May 29, 2022), 20 × 17 mm; follow-up image (June 3, 2023), 20 × 17 mm; and final image (March 27, 2024), 20 × 17 mm.

GPT-4o. Additionally, during this study, we did not design additional experiments for multiple nodules on a single CT image. Pathological examinations usually involve the tissue of a certain lesion; however, other lesions may produce different pathological results because of the existence of tumor heterogeneity.[38] Furthermore, a mean of 2.8 (SD, 1.2; range 2–10) CT examinations were found in this study. Future research should include longer follow-up periods and incorporate clinical records and biochemical tests to allow multimodal assessments. Moreover, we found that Molmo-7B also had the capability for lung nodule characterization in CT videos. However, limitations in multi-modal output currently restrict our understanding of its specific analytical strategies. Future evaluation of large multi-modal models with more parameters, such as Llama 3.2-90B and Molmo-72B, should be further considered. Finally, many DL models have been proposed for lung cancer classification.[7,8,15,32] In this study, we compared the performance of GPT-4o using two DL models, one local model and one previously published online model, as representatives. Future research should further compare GPT-4o with a broader range of DL models to strengthen the comparative evidence between LLMs and DL approaches.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Jiangdian Song (song.jd0910@gmail.com).

Materials availability

This study did not generate new, unique material or reagents.

Data and code availability

- Data: The lung CT videos, patient information, and results of the NLST dataset and the local training dataset have been uploaded to Mendeley Data: Mendeley Data, https://doi.org/10.17632/ggt4f4dr85.2. The pretrained GPT-4o was available at https://chatgpt.com/share/673dc28d-d3a0-8003-a186-acf5feb1b493. Note that this link may not be available due to updates made by ChatGPT; therefore, Video S1, Video S2, and Video S3 show the detailed training and testing process.
- This paper does not report original code. The software used in this study is described in the aforementioned section and the key resources table in detail.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors would like to thank all the radiologists and oncologists who participated in this study. This work was supported by the National Natural Science Foundation of China (92259104) and Scientific Research Fund of Liaoning Provincial Education Department (LJ242410159058).

AUTHOR CONTRIBUTIONS

J.S., L.Z., Y.M., N.X., and Y.W. contributed to the conceptual design of the study. N.X., Y.W., L.W., Y.M., and H.W. accessed and verified the raw data. N.X., Y.W., Y.M., and L.W. performed the statistical analysis. N.X., Y.W., L.W., and H.W. drafted the manuscript. J.S., Y.M., and L.Z. reviewed and revised the manuscript. All authors contributed to the interpretation of the data, had full access to all study data, and approved the final manuscript. All authors read and approved the final manuscript for submission.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANTS DETAILS
- METHOD DETAILS
  - CT images preprocessing and videos preparation
  - GPT-4o's memory preparation
  - Evaluation of the video of CT images using GPT-4o
  - Evaluation of the output of GPT-4o
  - Evaluation of changes in the nodule size by GPT-4o
  - Comparison of the performance of GPT-4o and Molmo-7B and claude
  - Test-retest of the reproducibility of evaluations
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2025.101988.

Received: September 3, 2024
Revised: November 21, 2024
Accepted: February 4, 2025
Published: March 4, 2025

REFERENCES

1. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. Nat. Med. *29*, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
2. Haver, H.L., Ambinder, E.B., Bahl, M., Oluyemi, E.T., Jeudy, J., and Yi, P.H. (2023). Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology *307*, e230424. https://doi.org/10.1148/radiol.230424.
3. Krishna, S., Bhambra, N., Bleakney, R., and Bhayana, R. (2024). Evaluation of Reliability, Repeatability, Robustness, and Confidence of GPT-3.5 and GPT-4 on a Radiology Board–style Examination. Radiology *311*, e232715. https://doi.org/10.1148/radiol.232715.
4. Sandmann, S., Riepenhausen, S., Plagwitz, L., and Varghese, J. (2024). Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat. Commun. *15*, 2050. https://doi.org/10.1038/s41467-024-46411-8.
5. Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A., and Zou, J. (2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell *186*, 1772–1791. https://doi.org/10.1016/j.cell.2023.01.035.
6. Wang, C., Shao, J., He, Y., Wu, J., Liu, X., Yang, L., Wei, Y., Zhou, X.S., Zhan, Y., Shi, F., et al. (2024). Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. Nat. Med. *30*, 3184–3195. https://doi.org/10.1038/s41591-024-03211-3.
7. Venkadesh, K.V., Setio, A.A.A., Schreuder, A., Scholten, E.T., Chung, K., W. Wille, M.M., Saghir, Z., van Ginneken, B., Prokop, M., and Jacobs, C. (2021). Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. Radiology *300*, 438–447. https://doi.org/10.1148/radiol.2021204433.
8. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer

screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. *25*, 954–961. https://doi.org/10.1038/s41591-019-0447-x.

9. Zhang, R., Wei, Y., Shi, F., Ren, J., Zhou, Q., Li, W., and Chen, B. (2022). The diagnostic and prognostic value of radiomics and deep learning technologies for patients with solid pulmonary nodules in chest CT images. BMC Cancer *22*, 1118. https://doi.org/10.1186/s12885-022-10224-z.
10. Ladbury, C., Amini, A., Govindarajan, A., Mambetsariev, I., Raz, D.J., Massarelli, E., Williams, T., Rodin, A., and Salgia, R. (2023). Integration of artificial intelligence in lung cancer: Rise of the machine. Cell Rep. Med. *4*, 100933. https://doi.org/10.1016/j.xcrm.2023.100933.
11. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.
12. Fink, M.A., Bischoff, A., Fink, C.A., Moll, M., Kroschke, J., Dulz, L., Heußel, C.P., Kauczor, H.-U., and Weber, T.F. (2023). Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology *308*, e231362. https://doi.org/10.1148/radiol.231362.
13. Huang, J., Yang, D.M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., Klesse, L.J., et al. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit. Med. *7*, 106. https://doi.org/10.1038/s41746-024-01079-8.
14. Zhang, C., Xu, J., Tang, R., Yang, J., Wang, W., Yu, X., and Shi, S. (2023). Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment. J. Hematol. Oncol. *16*, 114. https://doi.org/10.1186/s13045-023-01514-5.
15. González Maldonado, S., Delorme, S., Hüsing, A., Motsch, E., Kauczor, H.-U., Heussel, C.-P., and Kaaks, R. (2020). Evaluation of prediction models for identifying malignancy in pulmonary nodules detected via low-dose computed tomography. JAMA Netw. Open *3*, e1921221. https://doi.org/10.1001/jamanetworkopen.2019.21221.
16. Tammemagi, M., Ritchie, A.J., Atkar-Khattra, S., Dougherty, B., Sanghera, C., Mayo, J.R., Yuan, R., Manos, D., McWilliams, A.M., Schmidt, H., et al. (2019). Predicting malignancy risk of screen-detected lung nodules–mean diameter or volume. J. Thorac. Oncol. *14*, 203–211. https://doi.org/10.1016/j.jtho.2018.10.006.
17. Bueno, J., Landeras, L., and Chung, J.H. (2018). Updated Fleischner Society guidelines for managing incidental pulmonary nodules: common questions and challenging scenarios. Radiographics *38*, 1337–1350. https://doi.org/10.1148/rg.2018180017.
18. Prosper, A.E., Kammer, M.N., Maldonado, F., Aberle, D.R., and Hsu, W. (2023). Expanding role of advanced image analysis in CT-detected indeterminate pulmonary nodules and early lung cancer characterization. Radiology *309*, e222904. https://doi.org/10.1148/radiol.222904.
19. Hammer, M.M., Palazzo, L.L., Eckel, A.L., Barbosa, E.M., Jr., and Kong, C.Y. (2019). A decision analysis of follow-up and treatment algorithms for nonsolid pulmonary nodules. Radiology *290*, 506–513. https://doi.org/10.1148/radiol.2018180867.
20. Rami-Porta, R., Call, S., Dooms, C., Obiols, C., Sánchez, M., Travis, W.D., and Vollmer, I. (2018). Lung cancer staging: a concise update. Eur. Respir. J. *51*, 1800190. https://doi.org/10.1183/13993003.00190-2018.
21. Tsay, J.-C.J., Wu, B.G., Sulaiman, I., Gershner, K., Schluger, R., Li, Y., Yie, T.-A., Meyn, P., Olsen, E., Perez, L., et al. (2021). Lower airway dysbiosis affects lung cancer progression. Cancer Discov. *11*, 293–307. https://doi.org/10.1158/2159-8290.CD-20-0263.
22. Molmo 7B-D. https://huggingface.co/allenai/Molmo-7B-D-0924.
23. Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognit. Lett. *30*, 27–38. https://doi.org/10.1016/j.patrec.2008.08.010.
24. Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., and Dougherty, E.R. (2010). Small-sample precision of ROC-related estimates. Bioinformatics *26*, 822–830. https://doi.org/10.1093/bioinformatics/btq037.
25. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567. https://doi.org/10.1038/s41591-018-0177-5.
26. Hammer, M.M., Eckel, A.L., Palazzo, L.L., and Kong, C.Y. (2021). Cost-effectiveness of treatment thresholds for subsolid pulmonary nodules in CT lung cancer screening. Radiology *300*, 586–593. https://doi.org/10.1148/radiol.2021204418.
27. Mets, O.M., Schaefer-Prokop, C.M., and de Jong, P.A. (2018). Cyst-related primary lung malignancies: an important and relatively unknown imaging appearance of (early) lung cancer. Eur. Respir. Rev. *27*, 180079. https://doi.org/10.1183/16000617.0079-2018.
28. Miller, F.H., Lopes Vendrami, C., Recht, H.S., Wood, C.G., Mittal, P., Keswani, R.N., Gabriel, H., Borhani, A.A., Nikolaidis, P., and Hammond, N.A. (2022). Pancreatic cystic lesions and malignancy: assessment, guidelines, and the field defect. Radiographics *42*, 87–105. https://doi.org/10.1148/rg.210056.
29. Adams, S.J., Stone, E., Baldwin, D.R., Vliegenthart, R., Lee, P., and Fintelmann, F.J. (2023). Lung cancer screening. Lancet *401*, 390–408. https://doi.org/10.1016/S0140-6736(22)01694-4.
30. Bartlett, E.C., Silva, M., Callister, M.E., and Devaraj, A. (2021). False-negative results in lung cancer screening—evidence and controversies. J. Thorac. Oncol. *16*, 912–921. https://doi.org/10.1016/j.jtho.2021.01.1607.
31. Shao, X., Niu, R., Shao, X., Gao, J., Shi, Y., Jiang, Z., and Wang, Y. (2021). Application of dual-stream 3D convolutional neural network based on 18 F-FDG PET/CT in distinguishing benign and invasive adenocarcinoma in ground-glass lung nodules. EJNMMI Phys. *8*, 1–13. https://doi.org/10.1186/s40658-021-00423-1.
32. Cao, H., Liu, H., Song, E., Ma, G., Xu, X., Jin, R., Liu, T., and Hung, C.-C. (2020). A two-stage convolutional neural networks for lung nodule detection. IEEE J. Biomed. Health Inform. *24*, 2006–2015. https://doi.org/10.1109/JBHI.2019.2963720.
33. Sun, L., Zhang, M., Lu, Y., Zhu, W., Yi, Y., and Yan, F. (2024). Nodule-CLIP: Lung nodule classification based on multi-modal contrastive learning. Comput. Biol. Med. *175*, 108505. https://doi.org/10.1016/j.compbiomed.2024.108505.
34. Kim, R.Y., Oke, J.L., Pickup, L.C., Munden, R.F., Dotson, T.L., Bellinger, C.R., Cohen, A., Simoff, M.J., Massion, P.P., Filippini, C., et al. (2022). Artificial intelligence tool for assessment of indeterminate pulmonary nodules detected with CT. Radiology *304*, 683–691. https://doi.org/10.1148/radiol.212182.
35. Mazzone, P.J., and Lam, L. (2022). Evaluating the patient with a pulmonary nodule: a review. JAMA *327*, 264–273. https://doi.org/10.1001/jama.2021.24287.
36. Osarogiagbon, R.U., Liao, W., Faris, N.R., Fehnel, C., Goss, J., Shepherd, C.J., Qureshi, T., Matthews, A.T., Smeltzer, M.P., and Pinsky, P.F. (2023). Evaluation of lung cancer risk among persons undergoing screening or guideline-concordant monitoring of lung nodules in the Mississippi Delta. JAMA Netw. Open *6*, e230787. https://doi.org/10.1001/jamanetworkopen.2023.0787.
37. Dudurych, I., Pelgrim, G.-J., Sidorenkov, G., Garcia-Uceda, A., Petersen, J., Slebos, D.-J., de Bock, G.H., van den Berge, M., de Bruijne, M., and Vliegenthart, R. (2024). Low-dose CT–derived bronchial parameters in individuals with healthy lungs. Radiology *311*, e232677. https://doi.org/10.1148/radiol.232677.
38. Schneider, F., and Dacic, S. (2017). Histopathologic and molecular approach to staging of multiple lung nodules. Transl. Lung Cancer Res. *6*, 540–549. https://doi.org/10.21037/tlcr.2017.06.11.
39. Li, Q., Zhu, L., von Stackelberg, O., Triphan, S.M.F., Biederer, J., Weinheimer, O., Eichinger, M., Vogelmeier, C.F., Jörres, R.A., Kauczor, H.-U., et al. (2023). MRI compared with low-dose CT for incidental lung nodule detection in COPD: a multicenter trial. Radiol. Cardiothorac. Imaging *5*, e220176. https://doi.org/10.1148/ryct.220176.

40. Team, N.L.S.T.R. (2013). Data from the National Lung Screening Trial (NLST). The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.HMQ8-J677.

41. Wang, P., Xiao, X., Glissen Brown, J.R., Berzin, T.M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., et al. (2018). Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat. Biomed. Eng. *2*, 741–748. https://doi.org/10.1038/s41551-018-0301-3.

42. MacMahon, H., Naidich, D.P., Goo, J.M., Lee, K.S., Leung, A.N.C., Mayo, J.R., Mehta, A.C., Ohno, Y., Powell, C.A., Prokop, M., et al. (2017). Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology *284*, 228–243. https://doi.org/10.1148/radiol.2017161659.

43. Bankier, A.A., MacMahon, H., Goo, J.M., Rubin, G.D., Schaefer-Prokop, C.M., and Naidich, D.P. (2017). Recommendations for measuring pulmonary nodules at CT: a statement from the Fleischner Society. Radiology *285*, 584–600. https://doi.org/10.1148/radiol.2017162894.

44. Gierada, D.S., Rydzak, C.E., Zei, M., and Rhea, L. (2020). Improved interobserver agreement on lung-RADS classification of solid nodules using semiautomated CT volumetry. Radiology *297*, 675–684. https://doi.org/10.1148/radiol.2020200302.

45. Chen, Y., Zhong, J., Wang, L., Shi, X., Lu, W., Li, J., Feng, J., Xia, Y., Chang, R., Fan, J., et al. (2022). Robustness of CT radiomics features: consistency within and between single-energy CT and dual-energy CT. Eur. Radiol. *32*, 5480–5490. https://doi.org/10.1007/s00330-022-08628-3.

46. Berenguer, R., Pastor-Juan, M.D.R., Canales-Vázquez, J., Castro-García, M., Villas, M.V., Mansilla Legorburo, F., and Sabater, S. (2018). Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology *288*, 407–415. https://doi.org/10.1148/radiol.2018172361.

47. Ger, R.B., Zhou, S., Chi, P.-C.M., Lee, H.J., Layman, R.R., Jones, A.K., Goff, D.L., Fuller, C.D., Howell, R.M., Li, H., et al. (2018). Comprehensive investigation on controlling for CT imaging variabilities in radiomics studies. Sci. Rep. *8*, 13047. https://doi.org/10.1038/s41598-018-31509-z.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
GPT-4o	OpenAI	https://chatgpt.com/
Molmo-7B	Ai2	https://huggingface.co/allenai/Molmo-7B-D-0924
Python 3.11	Python Software Foundation	https://www.python.org/
R 4.2.0	R project	https://www.r-project.org/
Pytorch 2.2.2	Pytorch	https://pytorch.org/
ITK-SNAP 3.8.0	University of Pennsylvania	http://www.itksnap.org/pmwiki/pmwiki.php
R package pROC 1.18.0	pROC	https://cran.r-project.org/package=pROC
R package irr 0.84.1	irr	https://cran.r-project.org/package=irr
R package ggplot2 3.5.1	ggplot2	https://cran.r-project.org/package=ggplot2
Deposited data		
NLST dataset	NLST Research Team	https://www.cancerimagingarchive.net/collection/nlst/

EXPERIMENTAL MODEL AND STUDY PARTICIPANTS DETAILS

The CT images of all enrolled patients were chronologically sequenced and compiled as individual patient videos. CT images and corresponding nodule information of 20 patients were randomly selected and used to generate prompts and establish the GPT-4o memory bank. Subsequently, each patient's follow-up CT video and corresponding prompts were input into GPT-4o to estimate the probability of malignancy and evaluate changes in the nodule size and characteristics. The outputs generated by GPT-4o were independently evaluated by a panel of six radiologists with varying levels of experience.

The enrolled patients were from Shengjing Hospital of China Medical University (referred to as C1) and the Fourth Affiliated Hospital of China Medical University (C2) who underwent CT evaluations between January 31, 2018 and May 31, 2024 (Figure 2). The inclusion criteria were as follows: at least two contrast-enhanced/plain thin-slice lung CT scans obtained with a minimum interval of 2 months; available radiology reports included the nodule size (maximal long-axis and perpendicular maximal short-axis measurements in the same plane)[39]; and a biopsy performed after the last follow-up CT evaluation. The exclusion criteria were as follows: previous surgical resection of the lung tissue; suspected lung cancer without pathological confirmation; and metastases from other organs.

Additionally, to validate the ability of GPT-4o to screen for lung cancer, a dataset from C1 was collected (LLCS). The inclusion criteria were as follows: age $\geq$18 years; at least two low-dose lung CT images obtained at the center with a minimum interval of 10 months; complete demographic data; and the nodule size on each CT image was measured by radiologists (maximal long-axis and perpendicular maximal short-axis measurements in the same plane). The exclusion criteria were as follows: no nodule found during the evaluation by the radiologist; prior cancer diagnosis or thoracic surgery; and insufficient image quality.

To further validate the performance of GPT-4o using an independent cohort, we included an external test dataset of 100 randomly selected patients from the National Lung Screening Trial (NLST; n = 26,722) and the percentage of positive participants was 24%, according to the NLST screening database.[40] Additionally, patients with complete records of the imaging features (margin, predominant attenuation), nodule size, demographic characteristics (gender, age, smoking history), and at least two low-dose lung CT images with an interval of 1 year, were included in the GPT-4o analysis.

For the in-house datasets, CT images were acquired from the following scanners: Philips IQon -Spectral CT, Philips iCT 256, Philips Brilliance iCT 256, NMS Neuviz 128 Siemens, and SOMATOM go. Top scanner (Siemens Healthineers). The patient was placed in a supine position, with arms raised, and the lung was scanned at the end of inhalation. The following parameters were used: tube voltage: 100 kV or 140 kV, tube current: care dose 4D, scanning layer thickness: 2 mm, reconstructed layer thickness: 2 mm, reconstructed layer spacing: 2 mm, and matrix: 512 × 512, FOV: 350 mm × 350 mm. The enhanced scan was performed using a double-barrel high-pressure syringe to inject 70–90 mL of the non-ionic contrast agent iopromide intravenously into the cubital vein. The injection speed is 2.5–3.0 mL/s, and arterial phase images are obtained after 30–40 s of injection. All the participants' CT images were reviewed by senior experts at our hospitals. Hence, the radiology reports were standardized and included the location and size of the lung nodules, as agreed upon by the experts. Finally, a total of 647 patients were

retrospectively enrolled. Of these patients, 278 were from the C1 dataset, 191 were from the C2 dataset, 100 were from the NLST dataset, and 78 were from the LLCS dataset. The patients included from the respective participating centers constituted the experimental groups of each dataset in this study.

This multicenter retrospective study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (IRB) and Ethics Committee of Shengjing Hospital of China Medical University and The Fourth Affiliated Hospital of China Medical University [IRB Approval Number: 2024PS1395K and 2024-KS-151]. All data were anonymized to protect patient privacy, and informed consent was waived, as the study involved retrospective analysis of existing data without any direct patient interaction.

METHOD DETAILS

CT images preprocessing and videos preparation

Each CT scan was compiled as a video by converting each CT slice into sequential frames. CT images were preprocessed with a window width of 2000 and window level of −500 to allow an optimal visualization size (512 × 512 × N; with N representing the number of slices). These were converted to a video format at 20 frames per second to capture temporal progression of the findings similar to that during evaluations performed by radiologists.[41] All videos were anonymized to protect the privacy of the patients.

GPT-4o's memory preparation

Twenty patients were randomly selected to create a memory of lung nodules and generate textual prompts for GPT-4o. The prompts (Figure S5A) included demographic information (age, gender, smoking history), CT image time stamps, the corresponding image resolution and slice thickness, and the definition of the malignancy score. To facilitate nodule tracking across multiple images, the center coordinate and size of the lung nodule on the initial CT image were supplied to GPT-4o. To more accurately determine the malignancy of lung nodules, we also provided additional information (Figure S5B).[42]

For detection of nodule boundaries, GPT-4o utilized the provided coordinates to automatically identify the lesion and measure the nodule. After locating the frame that displayed the largest nodule size, GPT-4o delineated the nodule boundaries and reported the measured nodule size. The specific prompts for detecting nodule boundaries and measuring nodule size are shown in Figures S5C and S5D, respectively. Video S1 demonstrates the training process of GPT-4o.

Evaluation of the video of CT images using GPT-4o

During this study, GPT-4o was prompted to perform the following three primary functions for each CT image: evaluate the probability of lung nodule malignancy; evaluate the nodule size (maximal long-axis and perpendicular maximal short-axis measurements in the same plane); and comprehensively characterize the nodule features. Figure S4 provides details of the output of GPT-4o. Video S2 provide detailed demonstrations of the GPT-4o testing process for a participant, including nodule boundary detection and nodule size measurement.

Specifically, GPT-4o output the likelihood of lung nodule malignancy (range, 1–100; accurate to a single digit) for each CT image. Additionally, GPT-4o reported any observed changes in the predefined nodular features as the supporting diagnostic evidence. For the C1 and C2 datasets, nodule features encompassing shape (regular or irregular), margin (characteristics of the nodule's margin and changes in the margin morphology), internal structure and density (increased or decreased internal solid components and air cavities), presence of small bronchi involvement, presence or changes in small vessels around the nodule, and changes in the structure surrounding the nodule (e.g., pleural retraction, and vascular convergence.). For the NLST datasets, nodule features including margin [spiculated (stellate), smooth, and poorly defined], and predominant attenuation (soft tissue, ground glass, mixed) based on the NLST dataset document.

Evaluation of the output of GPT-4o

(1) Comparison of the GPT-4o's diagnosis and pathological results

This study compared the estimates of the probability of malignancy determined by GPT-4o and the pathological results (benign or malignant) using the C1, C2, and NLST datasets. The estimates determined by GPT-4o using the first CT image for each patient were compared to those of the corresponding pathological findings. This evaluation was repeated for each subsequent follow-up CT image to assess the ability of GPT-4o to incorporate longitudinal CT data.

(2) Radiologists' evaluation of the features detected by GPT-4o

Six radiologists with varying levels of experience (5, 10, 15, 15, 20, and 25 years) independently reviewed each CT image from the C1 and C2 datasets. For each patient, the radiologists documented their diagnostic observations of changes in the predefined features of these datasets. Subsequently, the radiologists rated the consistency of their observations with the feature descriptions generated by GPT-4o using a 5-point Likert scale (1, completely incorrect; 2, more incorrect than correct; 3, equally correct and incorrect; 4, more correct than incorrect; and 5, completely correct).

(3) Evaluations of lung cancer probability by GPT-4o and radiologists

For the LLCS dataset without pathological confirmation, two radiologists independently assessed the likelihood of malignancy of lung nodules based on the Lung-RADS criteria. The average of their probability estimations served as the reference lung cancer label. The intraclass correlation coefficient (ICC) was used to compare the probability of lung nodule malignancy estimated by GPT-4o with the average probability determined by the radiologists.

Evaluation of changes in the nodule size by GPT-4o

Across all datasets (C1, C2, LLCS, and NLST), the nodule size measurements performed by the radiologists were used as the gold standard. To compare the assessments of the nodule size by GPT-4o and the radiologists, this study followed the established method of assessing interobserver reliability for nodule size measurements. Specifically, the maximal long-axis and short-axis measurements of each nodule were averaged (the following equation), and fractional values were rounded up to the nearest millimeter (mm).[43,44]

$$Average\ diameter = (Long - axis\ diameter + Short - axis\ diameter) / 2$$

Additionally, to further determine the ability of GPT-4o to accurately track nodule size changes over time, we compared the assessments of changes in the nodule size performed by GPT-4o with those performed by radiologists. This comparison was specifically performed for nodules that exhibited increases or decreases in size based on the slope of the fitted line of the average diameter of the nodules on follow-up images.

Comparison of the performance of GPT-4o and Molmo-7B and claude

To further compare the performance of GPT-4o, we evaluated the performance of Molmo-7B, a state-of-the-art open-source, multimodal vision-language model, on longitudinal CT scan videos of lung nodules in the NLST dataset. Molmo models are trained on PixMo, a dataset comprising one million highly curated image-text pairs, and demonstrate near state-of-the-art performance on 11 academic benchmarks, second only to GPT-4o.[21] We utilized the latest version, "Molmo-7B-D-0924," for our experiments. We deployed Molmo-7B locally (using NVIDIA RTX A6000, Python 3.11, and PyTorch with CUDA 12.1) and constructed a conversational interface using the Chainlit package, enabling video display and prompt input. We subsequently analyzed each patient's CT image videos by providing the videos and corresponding prompts (including the same information as those provided to GPT-4o) to the Molmo-7B model. In addition, to compare GPT-4o with LLMs based on textual input, Claude was used for further comparison on nodule size evaluation and malignancy estimation.

Test-retest of the reproducibility of evaluations

A test-retest experiment was conducted to assess the reproducibility of GPT-4o's outputs. All six participating radiologists reviewed GPT-4o's responses to 50 randomly selected participants from in-house datasets in two rounds, with a 3-day interval between assessments. The same prompt, as described in Figure S4, was used in both rounds. In each instance, qualitative features and quantitative malignancy probabilities for each nodule were recorded. All six radiologists independently evaluated the consistency of GPT-4o's responses across the two rounds. For qualitative features (shape, margin, internal structure and density, presence of small bronchi involvement, presence or changes in small vessels around the nodule, and changes in the structure surrounding the nodule), consistency was empirically determined by the radiologists (range 1–100), and the average score was used to gauge GPT-4o's reproducibility. For quantitative malignancy probabilities, consistency was defined as a deviation of no more than 10 percentage points between the two rounds of estimations.[45–47]

Additionally, to assess the acceptability of the workflow using the assistance of GPT-4o, we conducted an ad-hoc experiment within the context of radiological practice. All participating radiologists quantitatively evaluated the following three distinct diagnostic models using a scale from 1 to 100: (1) GPT-4o-assisted lung nodule diagnosis, which provided diagnostic predictions and supporting evidence; (2) a local DL model, which provided only the diagnostic prediction; and (3) one online DL model with a reported accuracy exceeding 95%.[7] Radiologists rated each model based on their willingness to use the model in the field of radiology, reliance on the information provided, perceived potential for harm, the extent of inappropriate content, and the extent of missing content.

QUANTIFICATION AND STATISTICAL ANALYSIS

To evaluate the accuracy of GPT-4o to estimate the probability of malignancy compared to that of the pathological results, the receiver-operating characteristic curve and area under the receiver-operating characteristic curve (AUC) were used. Median scores with interquartile ranges (IQRs) and mean scores with standard deviations (SDs) were used to evaluate agreement between the characterization of the nodule features by GPT-4o and the radiologists. The ICC was used to assess agreement between the estimations of the probability of malignancy of the LLCS dataset performed by GPT-4o and the radiologists.

The ICC and Pearson's correlation coefficient were calculated to quantify the accuracy of measurements of the nodule size performed by GPT-4o. A Bland-Altman plot was generated to illustrate agreement between the nodule sizes measured by

GPT-4o and the radiologists. Additionally, the average of all the radiologists' evaluations during the ad hoc experiment was used to quantify the potential to apply these three models in the field of radiology.

GPT-4o was accessed through the chat interface, which includes upload functionality, on the OpenAI official website and used for all experiments. Statistical analyses were conducted using R software (version 4.2.0; R Foundation for Statistical Computing). Statistical significance was set at $p < 0.05$ (two-sided). The open-access dataset of this study has been uploaded to Mendeley Data, https://doi.org/10.17632/ggt4f4dr85.2.