

# Medical multimodal multitask foundation model for lung cancer screening

Received: 27 August 2024

Accepted: 31 January 2025

Published online: 11 February 2025



Chuang Niu<sup>1</sup>, Qing Lyu<sup>2</sup>, Christopher D. Carothers<sup>1</sup>, Parisa Kaviani<sup>3</sup>, Josh Tan<sup>2</sup>, Pingkun Yan<sup>1</sup>, Mannudeep K. Kalra<sup>3</sup>✉, Christopher T. Whitlow<sup>2</sup>✉ & Ge Wang<sup>1</sup>✉

Lung cancer screening (LCS) reduces mortality and involves vast multimodal data such as text, tables, and images. Fully mining such big data requires multitasking; otherwise, occult but important features may be overlooked, adversely affecting clinical management and healthcare quality. Here we propose a medical multimodal-multitask foundation model (M3FM) for three-dimensional low-dose computed tomography (CT) LCS. After curating a multimodal multitask dataset of 49 clinical data types, 163,725 chest CT series, and 17 tasks involved in LCS, we develop a scalable multimodal question-answering model architecture for synergistic multimodal multitasking. M3FM consistently outperforms the state-of-the-art models, improving lung cancer risk and cardiovascular disease mortality risk prediction by up to 20% and 10% respectively. M3FM processes multiscale high-dimensional images, handles various combinations of multimodal data, identifies informative data elements, and adapts to out-of-distribution tasks with minimal data. In this work, we show that M3FM advances various LCS tasks through large-scale multimodal and multitask learning.

Lung cancer remains the leading cause of cancer-related deaths<sup>1</sup>. Lung cancer screening (LCS) with low-dose computed tomography (LDCT) reduces lung cancer mortality by 20% in comparison with two-dimensional (2D) chest radiography in the National Lung Screening Trial (NLST)<sup>2</sup> and by 24% in comparison with no screening in the NELSON trial<sup>3</sup>. However, LCS faces challenges, such as the low screening rate (<10%)<sup>4</sup>, high false-positive rate<sup>5</sup>, sub-optimal workflows due to inadequate patient management<sup>6–9</sup>, under-utilization of multimodal data<sup>10,11</sup>, and particularly, a global shortage of radiologists for providing LCS. Hence, there is an important and immediate need for multidisciplinary efforts to broadly, equitably, and optimally implement LCS for minimized lung cancer mortality<sup>12</sup>.

Artificial intelligence (AI) promises to improve the quality and efficiency of LCS significantly. In particular, there is a vast amount of multimodal data accumulated over the past years, including low-dose

computed tomography (LDCT) images, patient demographics, smoking history, disease history, family cancer history, pathological results, follow-up data, etc.<sup>2</sup> LCS involves multiple clinical tasks, including lung nodule detection and characterization, Lung-RADS-based patient follow-up management, lung cancer risk estimation, and significant incidental findings such as diagnosis of various pulmonary, cardiovascular, and chest abnormalities. Deep learning methods achieved promising results for LCS-related tasks. For example, a deep learning method was proposed for lung cancer detection and risk estimation with LDCT in an end-to-end manner<sup>13</sup>. Recently, the Sybil model<sup>14</sup> was developed for lung cancer risk prediction using a single LDCT scan. In several studies<sup>15,16</sup>, deep learning models were developed for cardiovascular diseases (CVD) risk prediction with LDCT from LCS. Although promising, these LCS-related AI models were developed with relatively small single-modality datasets for individual tasks, limiting their

<sup>1</sup>Department of Biomedical Engineering, School of Engineering, Biomedical Imaging Center, Center for Computational Innovations, Center for Biotechnology & Interdisciplinary Studies, Rensselaer Polytechnic Institute, 110 8th Street, Troy 12180 NY, USA. <sup>2</sup>Department of Radiology, Wake Forest University School of Medicine, Winston-Salem 27103 NC, USA. <sup>3</sup>Department of Radiology, Massachusetts General Hospital, Harvard Medical School, White 270-E, 55 Fruit Street, Boston 02114 MA, USA. ✉e-mail: [mkalra@mgh.harvard.edu](mailto:mkalra@mgh.harvard.edu); [cwhitlow@wakehealth.edu](mailto:cwhitlow@wakehealth.edu); [wangg6@rpi.edu](mailto:wangg6@rpi.edu)

performance and utility in the multitask LCS workflow. Additionally, **the training schemes of current lung cancer risk models<sup>33,34</sup> require costly bounding box annotations, which makes building large-scale training datasets infeasible.**

In the fast-evolving AI field, **foundation models (FMs)** have shown previously unseen abilities to understand diverse data types and execute many tasks in a unified fashion<sup>17</sup>. Large FMs have updated the state-of-the-art (SoTA) performance across a wide range of tasks, such as natural language processing<sup>18–20</sup>, computer vision<sup>21,22</sup>, and multimodal understanding<sup>23–25</sup>. Inspired by these breakthroughs, increasing efforts have been made to develop medical foundation models, such as **biomedical language models<sup>26–31</sup>, medical vision-language models<sup>32–34</sup>, and generalist medical AI models<sup>35–37</sup>**. However, none of the existing FMs can effectively perform a variety of LCS tasks by interpreting the multimodal clinical data associated with LCS, particularly three-dimensional (3D) LDCT scans. **This limitation is primarily due to challenges in data curation and model architecture.** First, there is a high bar to systematically curate large-scale multimodal multitask datasets obtained in the real-world LCS workflow. Extensive domain-specific expertise and efforts are required to query, preprocess, and align 3D medical images and diverse structured/unstructured text-based clinical data with LCS-related tasks. Second, there is no scalable and adaptable foundation model dedicated to effectively interpreting multimodal LCS data, especially 3D CT images at different scales, and effectively performing diverse LCS-related tasks. Due to the high dimensionality of volumetric CT images, existing efforts only used small 2D/3D convolutional neural networks<sup>38</sup> and small vision Transformer models<sup>39</sup> at affordable computational costs.

In this work, **we present an integrated and scalable data curation approach to align high-dimensional medical images with other clinical datasets for LCS-related tasks**, including 17 LCS workflow-related tasks and 49 data elements. **Then, we develop a Medical Multimodal Multitask Foundation Model (M3FM) that perceives multimodal data including 3D CT volumes and various other clinical data, and performs multiple tasks involved in the LCS workflow.** Figure 1 illustrates the M3FM architecture, along with its training and inference processes. Extensive experiments show that our M3FM outperforms the previous SoTA models through large-scale multimodal and multitask learning, with the ability to identify informative data elements and adapt to the out-of-distribution task with a small dataset.

## Results

### Multimodal multitask datasets

**Figure 2a shows the general data curation pipeline, including medical tasks definition, task-specific multimodal data collection, multimodal data processing and alignment, and multimodal question-answering (MQA) dataset construction.** We target 17 (sub-)tasks in the LCS process, including 5 tasks for lung nodule detection and characterization, 1 task for cardiovascular disease (CVD) diagnosis, 1 task for CVD mortality risk prediction, 1 task for lung cancer risk prediction over multiple years, 7 tasks for other chest abnormality exams, 1 task for COVID-19 detection, as well as 1 task for American College of Radiology (ACR) guidelines for Lung CT Screening Reporting and Data System (Lung-RADS) categorization. COVID-19 detection from CT is included since it remains a global threat<sup>40</sup> and was reported in the LCS radiology reports collected from Massachusetts General Hospital (MGH) and Wake Forest University School of Medicine (WFUSM). The ground-truth labels come from different information sources, including radiology reports, disease history, pathology test results, follow-up data, death reports, and laboratory test results as described in Fig. 3a.

To curate the multimodal datasets, multiple data sources were aligned, including volumetric CT scans, demographics, smoking history, disease history, cancer history, family cancer history, and other task-specific clinical data. Race and ethnicity of NLST data are self-reported by participants using standardized questionnaires provided during the

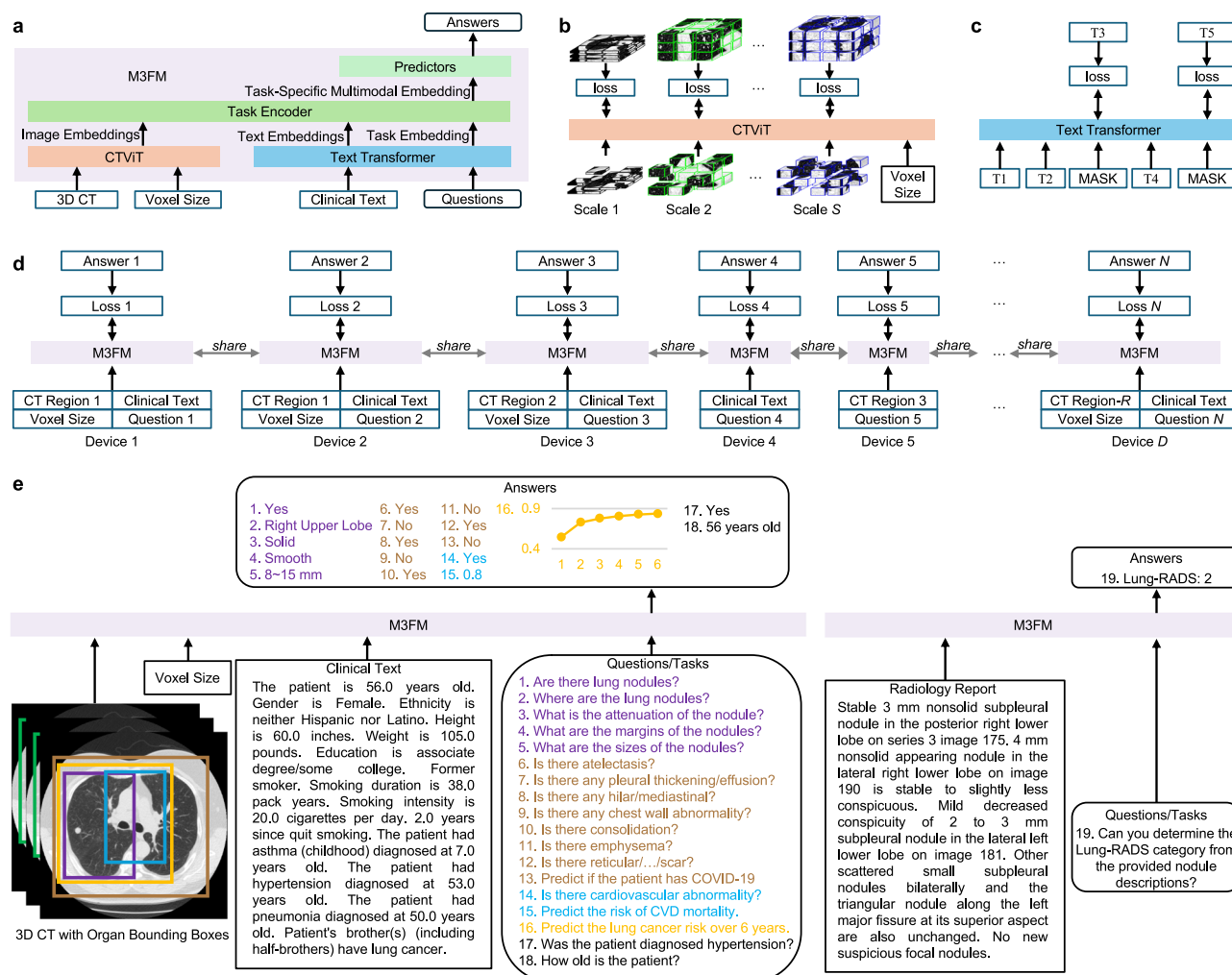
NLST enrollment process. In total, 49 different clinical data types were integrated into the multimodal datasets for LCS, as described in Fig. 3b. For each task, one training, one validation, and one or more testing datasets were constructed. Our multimodal multitask dataset is summarized in Fig. 2b. The data were collected from different data centers and institutes, including NLST, MIDRC, WFUSM, and MGH. In total, we curated 17 training, 17 validation, and 34 testing datasets for the 17 tasks, with detailed information in Fig. 2c–e. We also collected an out-of-distribution multimodal dataset from WFUSM for transfer learning. **To inspect the modeling ability for textural clinical data, we simulated a dataset for clinical information retrieval, as illustrated in Fig. 1e.** Since we unify multimodal multitask learning in an MQA framework, each dataset consists of task-specific multimodal inputs, questions, and answers. The details for all tasks are summarised in Fig. 3a.

As the first data source, we were granted access to all recorded data in NLST, **which is a randomized trial for evaluating LCS with 3D LDCT versus 2D chest radiography, demonstrating that screening with LDCT lowered lung cancer mortality by 20%.** The NLST data were collected from 33 medical institutions, which were randomly indexed without revealing their identifications publicly. The 26,722 participants in the LDCT screening arm were enrolled from August 2002 through April 2004. The participants underwent three screenings at 1-year intervals from August 2002 through September 2007. The follow-up data were collected until December 31, 2009. During the whole process, diverse data were recorded, including demographics, smoking history, disease history, multiple CT series with different reconstruction algorithms and associated imaging parameters, key abnormalities in fully structured reports, pathology test results for lung cancer, follow-up data, and vital status. Being consistent with the NLST clinical practice, we constructed 15 multimodal datasets for 15 tasks, including 5 datasets for predicting the presence of lung nodules and estimating the location, size, margin, and attenuation properties of lung nodules; 7 datasets for identifying chest abnormalities, including atelectasis, pleural thickening/effusion, non-calcified hilar/mediastinal adenopathy/mass, chest wall abnormality (bone destruction, metastasis, etc.), consolidation, emphysema, reticular/reticulonodular opacities/honeycombing/fibrosis/scar; 1 dataset for CVD diagnosis; 1 dataset for CVD mortality risk prediction following<sup>16</sup>, where the intervals between screening CT and CVD mortality are in the range from 11 days to 2619 days (within 8 years); and 1 dataset for lung cancer risk prediction within from 1 to 6 years as in<sup>14</sup>. For the CVD mortality risk prediction task, we further stratified the binary risk into 1–6 cut-off year risks following<sup>14</sup>. Each dataset was randomly split into training, validation, and test datasets. The patient information in the test dataset was not leaked to the training and validation datasets across all tasks. From NLST, we included 125,090 effective volumetric chest CT scans of the received 26,254 patient cases.

The second data source is the Medical Imaging and Data Resource Center (MIDRC)<sup>41</sup>, a collaboration of leading medical imaging organizations launched in August 2020 as part of NIBIB's response to the COVID-19 pandemic. We were granted to access all CT series with the associated clinical data. The ground-truth labels for COVID-19 were determined by either the Reverse Transcription Polymerase Chain Reaction (RT-PCR) or the Rapid Antigen Test (RAT). From MIDRC, we retrieved 35,730 volumetric chest CT series of 7609 patients scanned from 2011 to 2021. The patient data were randomly split into the training, validation, and test datasets.

All CT scans from NLST and MIDRC excluding those in any test datasets were combined as a CT pretraining dataset, comprised of 128,693 CT scans in total. To inspect if the clinical data are effectively encoded, we constructed a clinical question-answering dataset to retrieve key information from the textual clinical data. **The integration of all the above-curated datasets is called OpenM3Chest.**

To test the generalizability of M3FM, we independently collected two multimodal multitask datasets from the third and fourth data



**Fig. 1 | Overview of medical multimodal multitask foundation model (M3FM).**

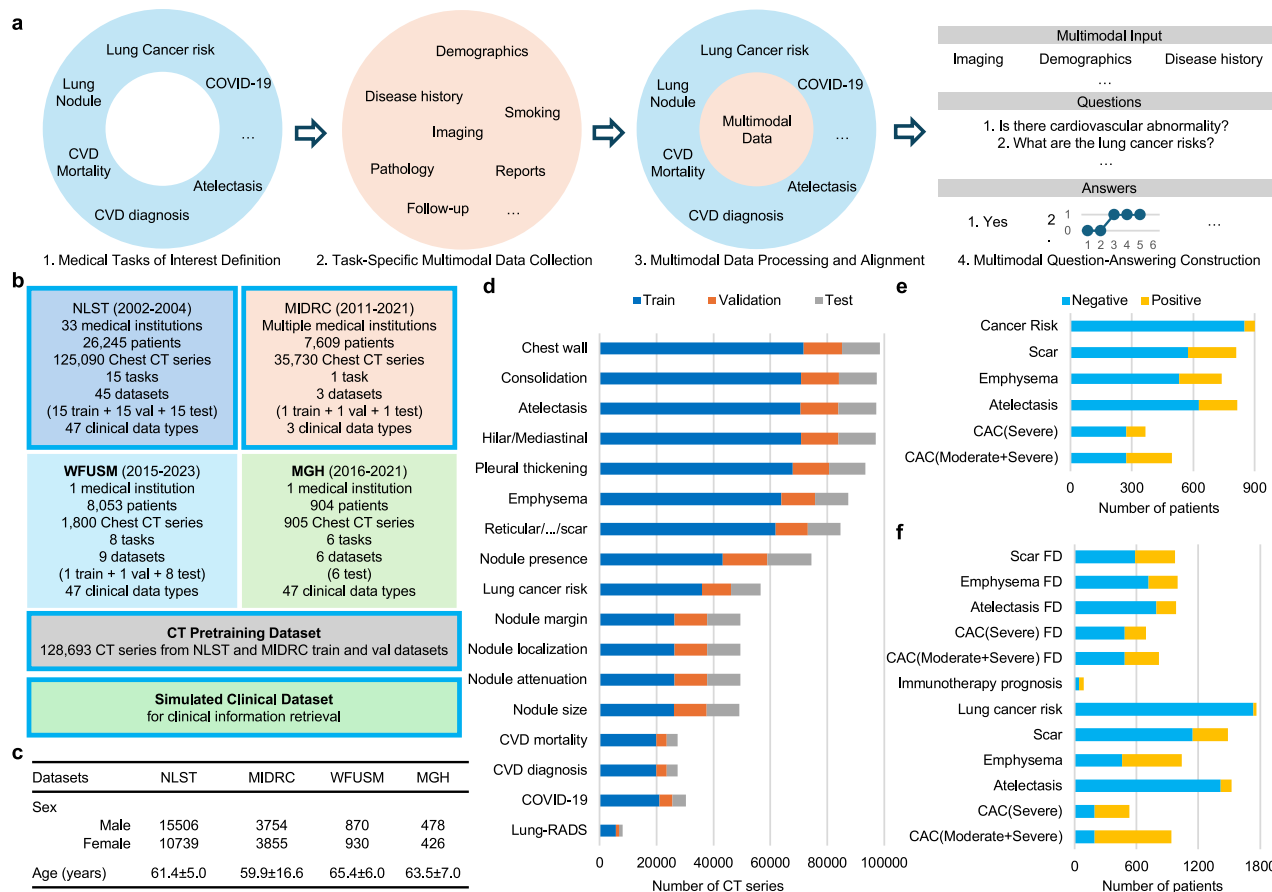
**a** M3FM architecture consists of four components: Computed Tomography Vision Transformer (CTViT), Text Transformer, Task Encoder, and Predictors. **b** Pretraining CTViT on multiscale CT volumes with voxel size-aware masked image modeling. Scale 1, Scale 2, ..., and Scale S denote S different sizes of images. **c** Pretraining Text Transformer with masked image models, where T1, T2, ..., and T5 denote a sequence of text tokens, T1, T2, and T4 are inputs, T3 and T5 are targets, and MASK is a special token meaning that the input token is masked. **d** Training the shared M3FM jointly with flexible multimodal and synergistic multitask learning using our distributed task-parallel strategy. Each device focuses on a single task with task-specific inputs, targets, and loss functions. D, N, and R denote the numbers of different devices, tasks, and image regions, respectively. Different tasks may have the same multimodal inputs on devices 1 and 2 and various multimodal or

single-modality inputs on devices 2, 3, 4, and 5. **e** M3FM inference flexibly handles multi-scale CT volumes (indicated by the rectangle boxes in different sizes and colors), clinical data, and multiple tasks. The colors of the CT bounding boxes match those of the questions and the predicted answers. For example, to answer Question 16, M3FM takes the orange region in the CT volume automatically localized using an organ localization model, the corresponding voxel size, and clinical text data as inputs. Questions 17 and 18 are two examples of auxiliary information retrieval tasks for clinical data modeling, which only take the clinical text as input. Question 19 predicts the Lung CT Screening Reporting and Data System (Lung-RADS) from lung nodule descriptions in a radiology report. reticular/.../scar reticular/reticulonodular opacities/honeycombing/fibrosis/scar, where / means or, COVID-19 Coronavirus Disease 2019.

sources, i.e., WFUSM and MGH, respectively. These multimodal LCS datasets include CT scans, radiology reports, demographics, smoking history, disease history, personal cancer history, family lung cancer history, and pathology test results for lung cancer. Race and ethnicity data of MGH and WFUSM datasets were collected from the MGH and WFUSM electronic health record systems and self-reported by participants. The radiology reports from WFUSM and MGH are in the structured reporting template with sub-headers, but the free text is used under each sub-header. We also collected a full-dose CT dataset with the associated radiology reports from WFUSM to evaluate the generalizability on full-dose CT scans. The MGH and WFUSM review boards approved the analysis of all these multimodal data and tasks. Based on the radiology reports and the pathology test results, we constructed 7 datasets from WFUSM and 6 datasets from MGH for independent evaluation, with the detailed information shown in

Fig. 2b, c, e, f. Specifically, we collected 8053 patient data from 2015 to 2023, all with radiology reports, and 1800 of them (from September 7, 2021 to December 30, 2022) with LDCT and multimodal information at WFUSM. We collected 1000 patient data with full-dose CT scans and the associated radiology reports from September 22, 2022, to December 31, 2022, at WFUSM. We collected 904 patient data with multimodal data at MGH from 2016 to 2021. The Lung-RADS dataset from WFUSM was randomly split into training, validation, and test datasets to classify the text descriptions into the Lung-RADS category. All other datasets of WFUSM and MGH were used for testing.

To evaluate the adaptability of our M3FM, we collected an out-of-distribution multimodal dataset for non-small cell lung cancer (NSCLC) immunotherapy prognosis from WFUSM. This dataset consists of 90 patient data, including the target label indicating if the patient was diagnosed with immune checkpoint-inhibitor-induced



**Fig. 2 | Dataset construction and summary.** **a** General data construction workflow consists of four steps: medical task definition, task-specific multimodal data collection, multimodal data processing and alignment, and multimodal question-answering construction. **b** The data used in this study was collected from two data centers, National Lung Screening Trial (NLST) and Medical Imaging and Data Resource Center (MIDRC), and two medical institutes, Wake Forest University School of Medicine (WFUSM) and Massachusetts General Hospital (MGH), with the key characteristics summarized, based on which a large volumetric Computed Tomography (CT) pretraining dataset and a simulated clinical dataset were constructed. The detailed configuration can be found in Supplementary Table 3. The

blue boxes indicate the OpenM3Chest dataset that is publicly available. **c** The patient sex and age distributions of the collected data from the involved data centers, where the age data represent mean age  $\pm$  standard deviation. **d** Distributions of the training, validation, and test datasets over all tasks. **e** Distributions of independent evaluation datasets from MGH. **f** Distributions of independent evaluation, full dose (FD) CT, and fine-tuning datasets from WFUSM. CVD Cardiovascular Disease, Reticular/.../scar, reticular/reticulonodular opacities/honeycombing/fibrosis/scar where / means or, COVID-19 Coronavirus Disease 2019, Lung-RADS Lung CT Screening Reporting and Data System, CAC Coronary Artery Calcification. Source data are provided as a Source Data file.

pneumonitis after immunotherapy, the CT scans before immunotherapy, and the clinical variables including the total cycles of Immuno-Oncology (IO), smoking information of pack years, Body Mass Index (BMI) at diagnosis, age, and if the patient received radiation prior to immunotherapy. Among the 90 patients, 49 patients developed immune checkpoint-inhibitor-induced pneumonitis, and the other patients were used as the control group.

Further details on the multimodal data processing and alignment and the MQA dataset construction are described in the Methods section.

### M3FM performance

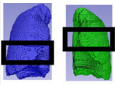

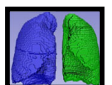
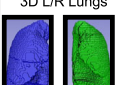
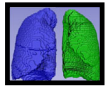
Figure 4a summarizes the key results of M3FM against the previous SOTA models<sup>14,16,34,42–45</sup> and the most powerful generalist AI model GPT-4o<sup>46</sup> on the OpenM3Chest dataset. The competing models are summarized in Supplementary Table 1. We used the Area Under the receiver operating characteristic Curve (AUC) and the 95% two-sided Confidence Intervals (CI) of AUC as the evaluation metrics<sup>47</sup>.

With the detailed comparative results summarized in Supplementary Table 2, M3FM outperformed the previous SOTA models across all tasks, demonstrating significant improvements in most of

them. Specifically, for a fair comparison, we retrained the Sybil model<sup>14</sup>, denoted as Sybil', for lung cancer risk prediction without using costly bounding box annotations but predicting lung cancer risks by merging the separate results of left and right lungs. It is observed that Sybil' achieved inferior results for 1 - 2-year risk prediction but superior results for 3 - 6-year risk prediction in comparison with the results obtained using the original Sybil model. Without using any bounding box, our M3FM achieved an AUC of 0.9400 (95% Confidence Intervals = 0.9119–0.9698), 0.8881 (95% Confidence Intervals = 0.8567–0.9195), 0.8599 (95% Confidence Intervals = 0.8288–0.8910), 0.8604 (95% Confidence Intervals = 0.8310–0.8898), 0.8392 (95% Confidence Intervals = 0.8098–0.8685), 0.8232 (95% Confidence Intervals = 0.7936–0.8529) for lung cancer risk prediction over six years, outperforming both Sybil' and original Sybil models by the margins of 5% to 9% and 2% to 11%, respectively. For CVD diagnosis and CVD mortality prediction, we compared the results on both the original dataset<sup>16</sup> and our OpenM3Chest dataset. M3FM achieved an AUC of 0.9284 (95% Confidence Intervals = 0.9136–0.9433) for CVD diagnosis and an AUC of 0.8904 (95% Confidence Intervals = 0.8427–0.9381) for CVD mortality prediction on the OpenChest dataset, outperforming the previous model (Tri2D-Net<sup>16</sup>) by 5% and 9%



a

Text Input	Image Input	Example Questions	Candidate Answers	Label Source
Demographics		Is there any lung nodule?	A: Yes; B: No	Radiology report
		Where is the lung nodule?	A: Right upper lobe; B: Right middle lobe; C: Right lower lobe; D: Left upper lobe; E: Left lower lobe	
		What is the attenuation of the nodule?	A: Solid; B: Ground Glass; C: Others(Part-Solid, Fluid/Water, Fat, Undetermined)	
		What is the type of nodule margin?	A: Spiculated (Stellate); B: Smooth; C: Poorly defined; D: Unable to determine	
Smoking History		What is the size of the nodule?	A: <=4mm; B: 4~6mm; C: 6~8mm; D: 8~15mm; E: 15~30mm; F: >30mm	Radiology report, disease history, death report
Disease History		Predict the risk of cardiovascular disease mortality.	Risk value: 0~1	
	Is there significant cardiovascular abnormality?	A: Yes; B: No	Radiology report	
	Is there atelectasis?			
	Is there pleural thickening/effusion?			
Is there hilar/mediastinal adenopathy/mass?				
Is there chest wall abnormality?				
Is there consolidation?				
Cancer History		Is there emphysema?	Is there any reticular/reticulonodular opacities/honeycombing/fibrosis/scar?	
		Family History		
2-year risk value: 0~1				
3-year risk value: 0~1				
4-year risk value: 0~1				
5-year risk value: 0~1				
6-year risk value: 0~1				
Demographics		Predict if the patient has COVID-19 given the data.	A: Yes; B: No	RT-PCR or RAT
Clinical data prior to immunotherapy		Will the immunotherapy induce pneumonitis?		
Nodule Report	N/A	What is the Lung-RADS?	A: 1; B: 2; C: 3; D: 4	Predefined
Simulated Clinical Data	N/A	Was the patient diagnosed diabetes? How old is the patient? ...	Yes, No, or Number	

b

Multimodal Data Elements	
Demographics	Age
	Gender
	Race
	Ethnic
Smoking History	Height
	Weight
	Education
	Smoking status
Disease History & Diagnosis Age	Package years
	Smoke day
	Age quit
	Asthma(adult)
	Asbestosis
	Bronchiectasis
	Asthma(childhood)
	Chronic bronchitis
	COPD
	Diabetes
	Emphysema
	Lung Fibrosis
	Heart disease/attack
	Hypertension
Personal Cancer History & Diagnosis Age	Pneumonia
	Sarcoidosis
	Silicosis
	Stroke
	Tuberculosis
	Bladder
	Breast
	Cervical
	Colorectal
	Esophageal
	Kidney
	Larynx
	Lung
	Nasal
Oral	
Family Lung Cancer History	Pancreatic
	Pharynx
	Stomach
	Thyroid
	Transitional Cell
	Brother
Immunotherapy clinical data	Child
	Father
	Mother
	Sister
CT	Cycles of IO
	Pack years
	BMI
	Age at diagnosis
	Radiation
	3D CT
	Voxel size

**Fig. 3 | Overview of the multimodal question-answering datasets.** **a** Alignment among text input, image input, example questions, and candidate answers. The black bounding boxes on lungs and heart illustrate input region sizes, including 2.5-dimensional left or right lung regions (2.5D L/R Lung), three-dimensional heart regions, three-dimensional left and right lung regions (3D L&R Lungs), three-

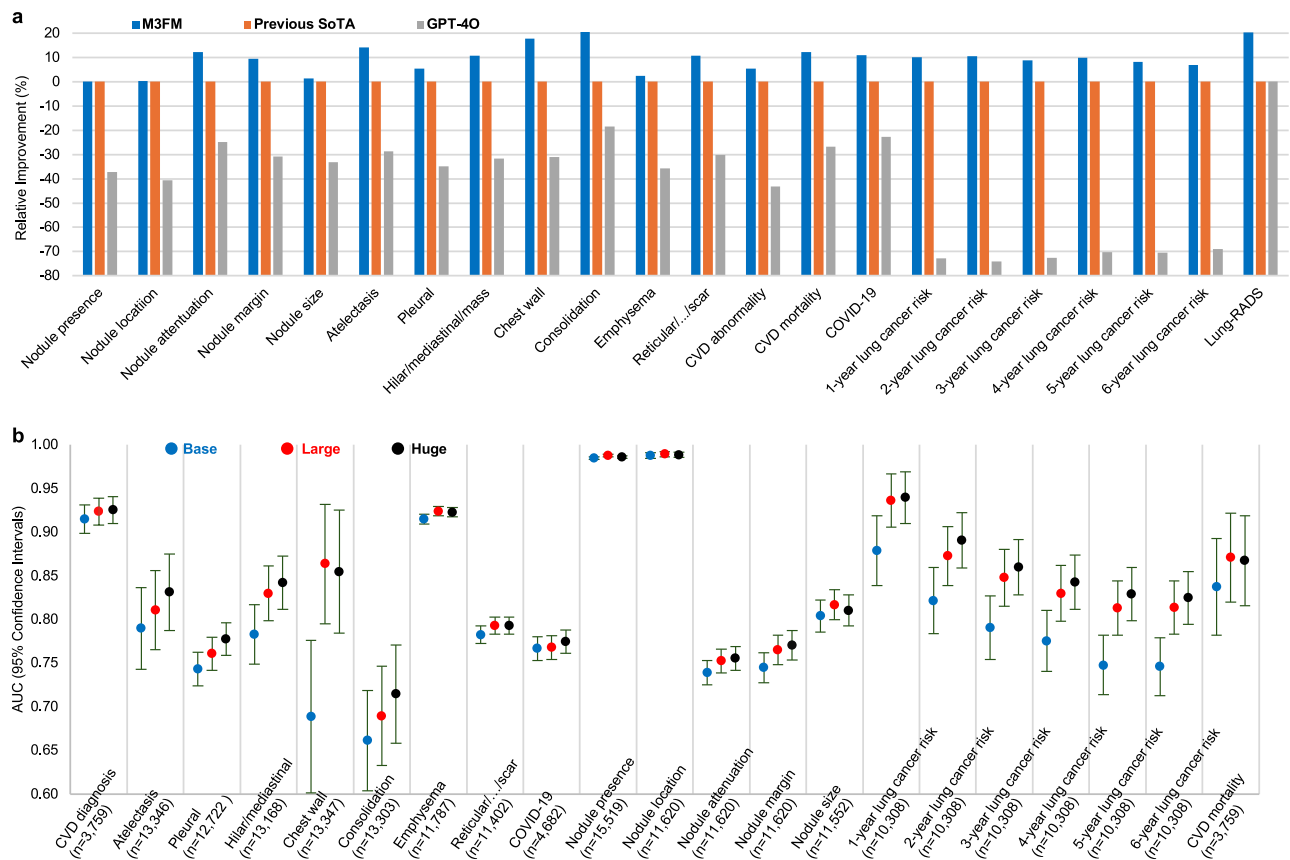
dimensional heart regions, three-dimensional left or right lung regions (3D L/R Lungs). **b** Multimodal data elements involved in this work including three-dimensional (3D) computed tomography (CT). Patient information on race and ethnicity is self-reported. COVID-19 Coronavirus Disease 2019, Lung-RADS Lung CT Screening Reporting and Data System.

respectively, and achieved an AUC of 0.9304 (95% Confidence Intervals = 0.9150–0.9458) for CVD diagnosis and 0.8606 (95% Confidence Intervals = 0.8063–0.9150) for CVD mortality prediction on the datasets constructed in<sup>16</sup>, outperforming the previous model (Tri2D-Net) by ~ 5% and ~ 10% respectively. On average, M3FM enhanced the 1-6 year CVD mortality risk prediction performance by 14.22% in AUC compared to the previous best model (see Supplementary Table 2). For several tasks including nodule detection, nodule localization, nodule size prediction, and emphysema detection, M3FM improved the results by various degrees up to 3% of AUC. For all the other tasks, M3FM significantly improved the performance from ~ 5% to ~ 10%. To study the scalability of M3FM, we trained three versions of M3FM, consisting of 257M (M3FM-Base), 502M (M3FM-Large), and 865M (M3FM-Huge) trainable parameters respectively. The results obtained using these three models are summarized in Fig. 4b. Overall, with a larger model size, the performance became better, especially from M3FM-Base to M3FM-Large. This trend is consistent with the well-known scaling law<sup>48</sup> in the field of foundation models.

**M3FM encoding multimodal data and synergizing multiple clinical tasks**

Table 1 compares the results of the single-modality single-task, multi-modality single-task, and multi-modality multitask M3FM-Large models. First, the single-modality single-task models were trained and evaluated on LDCT data only and denoted by M3FM-SM-ST, while the multi-modality single-task models were trained and evaluated on multimodal data and denoted by M3FM-MM-ST. Overall, the

multimodal information improved the prediction results for multiple tasks. In particular, M3FM-SM-ST achieved an AUC of 0.8163 (95% Confidence Intervals = 0.7585–0.8741) for CVD mortality prediction while the M3FM-MM-ST model achieved an AUC of 0.8709 (95% Confidence Intervals = 0.8200–0.9219), which represents a 5.46% improvement. Similarly, for multi-year CVD mortality risk prediction, the multimodal model outperformed the single-modality model by 5% on average as shown in Supplementary Table 2. While M3FM-SM-ST achieved an AUC of 0.8924 (95% Confidence Intervals = 0.8745–0.9104) for CVD diagnosis, the M3FM-MM-ST model achieved an AUC of 0.9238 (95% Confidence Intervals = 0.9084–0.9392), i.e., a 3.14% improvement. Similarly, M3FM-SM-ST achieved an AUC of 0.6515 (95% Confidence Intervals = 0.5939–0.7092) for consolidation detection, and the M3FM-MM-ST model achieved an AUC of 0.6895 (95% Confidence Intervals = 0.6326–0.7464), a 3.80% improvement. Also, M3FM-SM-ST achieved an AUC of 0.7676 (95% Confidence Intervals = 0.7573–0.7779) for reticular/reticulonodular opacities/honeycombing/fibrosis/scar detection, and the M3FM-MM-ST model achieved an AUC of 0.7929 (95% Confidence Intervals = 0.7830–0.8027), a 2.53% improvement. It is further observed that M3FM-MM-ST models produce slightly improved or comparable results in comparison with M3FM-SM-ST for the other tasks. Then, we compared the multimodal multitask model (M3FM-MM-MT) and multimodal single-task models (M3FM-MM-ST). Impressively, training on multiple tasks, M3FM-MT-MM outperformed the M3FM-ST-MM for 17 out of 22 (sub)-tasks. In reference to the label distributions of the multiple tasks in Supplementary Table 3, the five tasks that were not benefited from multitask



**Fig. 4 | Overall and Scalable performance of the Medical Multimodal Multitask Foundation Model (M3FM).** **a** Comparison of the best M3FMs with previous state-of-the-art (SoTA) models in including Generative Pre-trained Transformer 4 Omni (GPT-4O) in terms of Area Under the Curve (AUC) relative improvement. The compared models have been summarized in Supplementary Table 1. The AUC values and 95% confidence intervals of all models can be found in Supplementary Table 2. **b** AUC results with 95% confidence intervals for M3FM models of three

scales including Base, Large, and Huge. The AUC value and two-sided 95% confidence interval for each task were calculated from its entire test dataset. Error bars in **b** indicate the two-sided 95% confidence intervals. CVD Cardiovascular Disease, Reticular/.../scar reticular/reticulonodular opacities/honeycombing/fibrosis/scar, where / means or, COVID-19 Coronavirus Disease 2019, Lung-RADS Lung CT Screening Reporting and Data System. Source data are provided as a Source Data file.

learning have the largest balance ratios of the number of minority class samples over the number of majority class samples. In other words, multitask learning is more beneficial for tasks with more imbalanced datasets or a much smaller number of positive/minority class labels.

### M3FM identifying clinically informational elements

Since M3FM accommodates any combination of multimodal datasets in the training and inference stages, we investigated the application of M3FM to analyze the synergy between multimodal data elements and clinical tasks by observing the effects of different input combinations on the model outcomes. Table 2 presents the ablation results using different combinations of multimodal data for CVD diagnosis and mortality prediction. M3FM using all multimodal inputs improved the AUC by 3% ~ 4% relative to the results using LDCT only and by 12% and 5% over that using clinical data only for CVD diagnosis and mortality prediction respectively. Furthermore, the M3FM results show that the disease histories of heart disease or heart attack, hypertension, stroke, and diabetes consistently boosted the AUC results by gradually adding them into the input combination for CVD diagnosis and mortality prediction. Supplementary Table 4 shows the lung cancer risk prediction results using different inputs, showing that demographic information slightly improved the AUC results.

Then, we evaluated if M3FMs could effectively encode the physical size information. The ablation results in Fig. 5a show that the embedded physical size of LDCT improved the AUC results for multiple tasks. The physical size information boosted the AUC of 1 - 6-year

lung cancer risk prediction by 5%, 4%, 4%, 7%, 8% and 12%, respectively. The physical size information also improved AUC results of the nodule size characterization, CVD diagnosis, and CVD mortality prediction, by 0.71%, 0.47%, and 1.11% respectively.

We quantitatively evaluated the relevance of different clinical elements with model outputs by visualizing the attention maps of the last task attention block in M3FM. Figure 5b visualizes the attention heat maps on selected CT slices and text tokens of individual patients with CVD or lung cancer risks. In CVD diagnosis, the coronary artery calcification areas are highlighted in the LDCT attention heat maps, and the patients' disease histories of diabetes, heart disease or heart attack, hypertension, and stroke are highly relevant among text tokens, which is consistent with the quantitative results in Table 2. Furthermore, the ablation inference in Supplementary Fig. 1 explicitly shows how the information from multiple sources is composed to affect the model prediction. In a case of positive CVD diagnosis, M3FM failed predictions when taking LDCT only or LDCT plus uninformative clinical data as inputs. The same M3FM successfully diagnosed CVD when using LDCT plus the relevant clinical data including the diabetes/heart disease history. This is consistent with the ablation results summarized in Table 2, where multimodal inputs are statistically beneficial for the M3FM inferences. In predicting lung cancer risks, the lung nodules in LDCT images are localized in the heat maps, and the text tokens related to demographic and family lung cancer histories are more correlated to the model outputs as shown in Fig. 5c. Although the visualization of the attention maps provides a window to inspect

**Table 1 | Comparison of M3FM Variants on the OpenM3Chest Dataset**

Tasks	M3FM-SM-ST	M3FM-MM-ST	M3FM-MM-MT
<b>Nodule presence</b>	<b>0.9877 (0.9863–0.9892)</b>	<b>0.9876 (0.9862–0.9891)</b>	<b>0.9858 (0.9843–0.9874)</b>
Nodule location (Average)	0.9884	0.9893	0.9877
Right Upper Lobe	0.9912 (0.9887–0.9937)	0.9915 (0.9891–0.9939)	0.9895 (0.9868–0.9922)
Right Middle Lobe	0.9811 (0.9760–0.9862)	0.9826 (0.9777–0.9875)	0.9793 (0.9739–0.9846)
Right Lower Lobe	0.9856 (0.9824–0.9887)	0.9865 (0.9835–0.9895)	0.9850 (0.9818–0.9882)
Left Upper Lobe	0.9918 (0.9892–0.9944)	0.9928 (0.9904–0.9952)	0.9922 (0.9896–0.9947)
Left Lower Lobe	0.9922 (0.9897–0.9946)	0.9932 (0.9909–0.9955)	0.9924 (0.9900–0.9948)
Nodule attenuation (Average)	0.7540	0.7525	0.7589
Solid	0.7817 (0.7728–0.7906)	0.7800 (0.7711–0.7889)	0.7888 (0.7802–0.7975)
Ground Glass	0.8410 (0.8285–0.8534)	0.8337 (0.8210–0.8463)	0.8533 (0.8413–0.8653)
Others	0.6393 (0.6195–0.6592)	0.6437 (0.6238–0.6636)	0.6345 (0.6146–0.6543)
Nodule margin (Average)	0.7637	0.7653	0.7742
Spiculated	0.7929 (0.7760–0.8097)	0.7893 (0.7724–0.8062)	0.8156 (0.7995–0.8317)
Smooth	0.7892 (0.7811–0.7974)	0.7805 (0.7722–0.7888)	0.7939 (0.7859–0.8020)
Poorly Defined	0.7750 (0.7628–0.7871)	0.7511 (0.7386–0.7636)	0.7652 (0.7529–0.7774)
Undetermined	0.6975 (0.6667–0.7284)	0.7404 (0.7105–0.7703)	0.7222 (0.6919–0.7524)
Nodule size (Average)	0.8230	0.8167	0.8195
≤4 mm	0.7732 (0.7624–0.7839)	0.7760 (0.7653–0.7867)	0.7794 (0.7688–0.7901)
4–6 mm	0.7006 (0.6908–0.7104)	0.6782 (0.6682–0.6882)	0.6976 (0.6878–0.7075)
6–8 mm	0.7542 (0.7410–0.7674)	0.7330 (0.7195–0.7465)	0.7278 (0.7142–0.7413)
8–15 mm	0.8681 (0.8556–0.8807)	0.8595 (0.8466–0.8723)	0.8660 (0.8535–0.8786)
15–30 mm	0.9102 (0.8921–0.9284)	0.9125 (0.8946–0.9305)	0.9159 (0.8982–0.9335)
> 30 mm	0.9316 (0.8914–0.9718)	0.9413 (0.9038–0.9789)	0.9302 (0.8897–0.9708)
CVD Abnormality	0.8924 (0.8745–0.9104)	0.9238 (0.9084–0.9392)	0.9284 (0.9136–0.9433)
CVD Mortality	0.8163 (0.7585–0.8741)	0.8709 (0.8200–0.9219)	0.8904 (0.8427–0.9381)
Atelectasis	0.8172 (0.7723–0.8622)	0.8108 (0.7654–0.8563)	0.8181 (0.7238–0.8200)
Pleural thickening/effusion	0.7373 (0.7178–0.7567)	0.7607 (0.7417–0.7797)	0.7657 (0.7468–0.7846)
Hilar/mediastinal adenopathy/mass	0.8299 (0.7985–0.8613)	0.8297 (0.7983–0.8611)	0.8328 (0.8017–0.8639)
Chest wall abnormality	0.8151 (0.6614–0.9689)	0.8239 (0.6859–0.9718)	0.8344 (0.6862–0.9826)
Consolidation	0.6515 (0.5939–0.7092)	0.6895 (0.6326–0.7464)	0.7241 (0.6683–0.7798)
Emphysema	0.9137 (0.9079–0.9194)	0.9240 (0.9186–0.9294)	0.9119 (0.9061–0.9177)
Reticular/reticulonodular opacities/ honeycombing/fibrosis/scar	0.7676 (0.7573–0.7779)	0.7929 (0.7830–0.8027)	0.7744 (0.7643–0.7846)
Lung Rads (Average)	0.8565	N/A	0.8706
1	0.8957 (0.8735–0.9179)	N/A	0.9036 (0.8831–0.9240)
2	0.8753 (0.8559–0.8946)	N/A	0.8733 (0.8539–0.8926)
3	0.8306 (0.7728–0.8884)	N/A	0.8335 (0.7761–0.8909)
4	0.8245 (0.7644–0.8846)	N/A	0.8720 (0.8186–0.9255)
1-Year Cancer Risk	0.9298 (0.8980–0.9615)	0.9362 (0.9058–0.9666)	0.9400 (0.9119–0.9698)
2-Year Cancer Risk	0.8697 (0.8358–0.9036)	0.8727 (0.8391–0.9063)	0.8881 (0.8567–0.9195)
3-Year Cancer Risk	0.8418 (0.8088–0.8748)	0.8479 (0.8154–0.8805)	0.8599 (0.8288–0.8910)
4-Year Cancer Risk	0.8338 (0.8020–0.8655)	0.8299 (0.7979–0.8619)	0.8604 (0.8310–0.8898)
5-Year Cancer Risk	0.8088 (0.7773–0.8402)	0.8131 (0.7819–0.8443)	0.8392 (0.8098–0.8685)
6-Year Cancer Risk	0.7922 (0.7606–0.8238)	0.8135 (0.7829–0.8440)	0.8232 (0.7936–0.8529)
COVID-19	0.7688 (0.7553–0.7823)	0.7679 (0.7544–0.7814)	0.7569 (0.7431–0.7706)

SM denotes single-modality, MM signifies multi-modality, ST represents single-task, and MT indicates multitask. Average denotes the mean AUC value of all sub-categories. AUC values with 95% confidence intervals in parentheses are reported. M3FM-MM-ST outperforms M3FM-SM-ST models in 14 of 21 tasks. M3FM-MM-MT outperforms M3FM-MM-ST models in 17 of 22 tasks. N/A means not available.

the behavior of the Transformer model, it is not always reliable to reveal the correlation between model predictions and input tokens, e.g., less relevant tokens were highlighted in Fig. 5b, which is consistent to the prior findings<sup>49,50</sup>.

**M3FM improving generalizability**

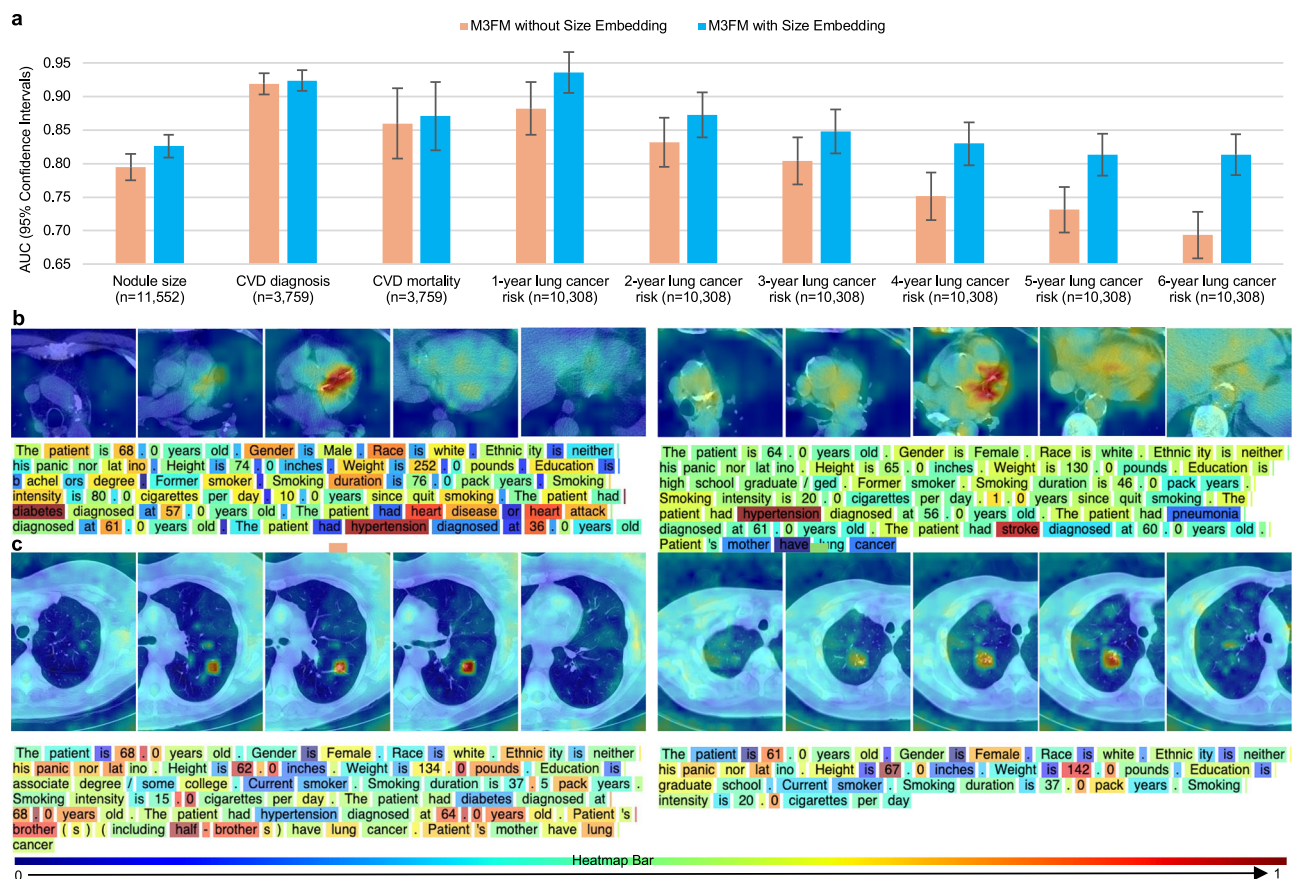
We evaluated the generalizability of M3FMs on the multimodal datasets independently collected from MGH and WFUSM, with the

comparative results shown in Fig. 6a, b, respectively. For the CVD diagnosis task, we constructed two datasets, which regard (1) moderate and severe CVD as positive and (2) severe CVD only as positive, respectively. On the two MGH CVD datasets, the multimodal multitask model (M3FM-MM-MT) improved the AUC by 10.60% and 6.57% relative to the previous model, improved the AUC by 4.85% and 2.36% relative to the single-modality single-task model (M3FM-SM-ST), and also achieved slight AUC improvements relative to the multi-modality

**Table 2 | Evaluation of clinical data elements in the CVD tasks**

Train input	Test input	CVD diagnosis	CVD mortality
LDCT	LDCT	0.8924 (0.8745–0.9104)	0.8163 (0.7585–0.8741)
LDCT + All clinical data	LDCT	0.8846 (0.8661–0.9030)	0.8344 (0.7786–0.8902)
LDCT + All clinical data	LDCT + All clinical data	0.9238 (0.9084–0.9392)	0.8709 (0.8200–0.9219)
LDCT + All clinical data	LDCT + All disease history	0.9237 (0.9083–0.9391)	0.8709 (0.8200–0.9218)
LDCT + All clinical data	LDCT + Heart disease/attack	0.9001 (0.8828–0.9175)	0.8327 (0.7767–0.8887)
LDCT + All clinical data	LDCT + Hypertension	0.9131 (0.8978–0.9294)	0.8641 (0.8122–0.9160)
LDCT + All clinical data	LDCT + Diabetes	0.8929 (0.8749–0.9108)	0.8473 (0.7930–0.9015)
LDCT + All clinical data	LDCT + Stroke	0.8897 (0.8715–0.9078)	0.8454 (0.7910–0.8999)
LDCT + All clinical data	LDCT + Heart disease/attack, Hypertension	0.9221 (0.9065–0.9377)	0.8689 (0.8177–0.9201)
LDCT + All clinical data	LDCT + Heart disease/attack, Hypertension, Stroke	0.9227 (0.9072–0.9382)	0.8684 (0.8171–0.9197)
LDCT + All clinical data	LDCT + Heart disease/attack, Stroke, Hypertension, Diabetes	0.9246 (0.9093–0.9400)	0.8729 (0.8223–0.9235)

AUC values with 95% confidence intervals in parentheses are reported



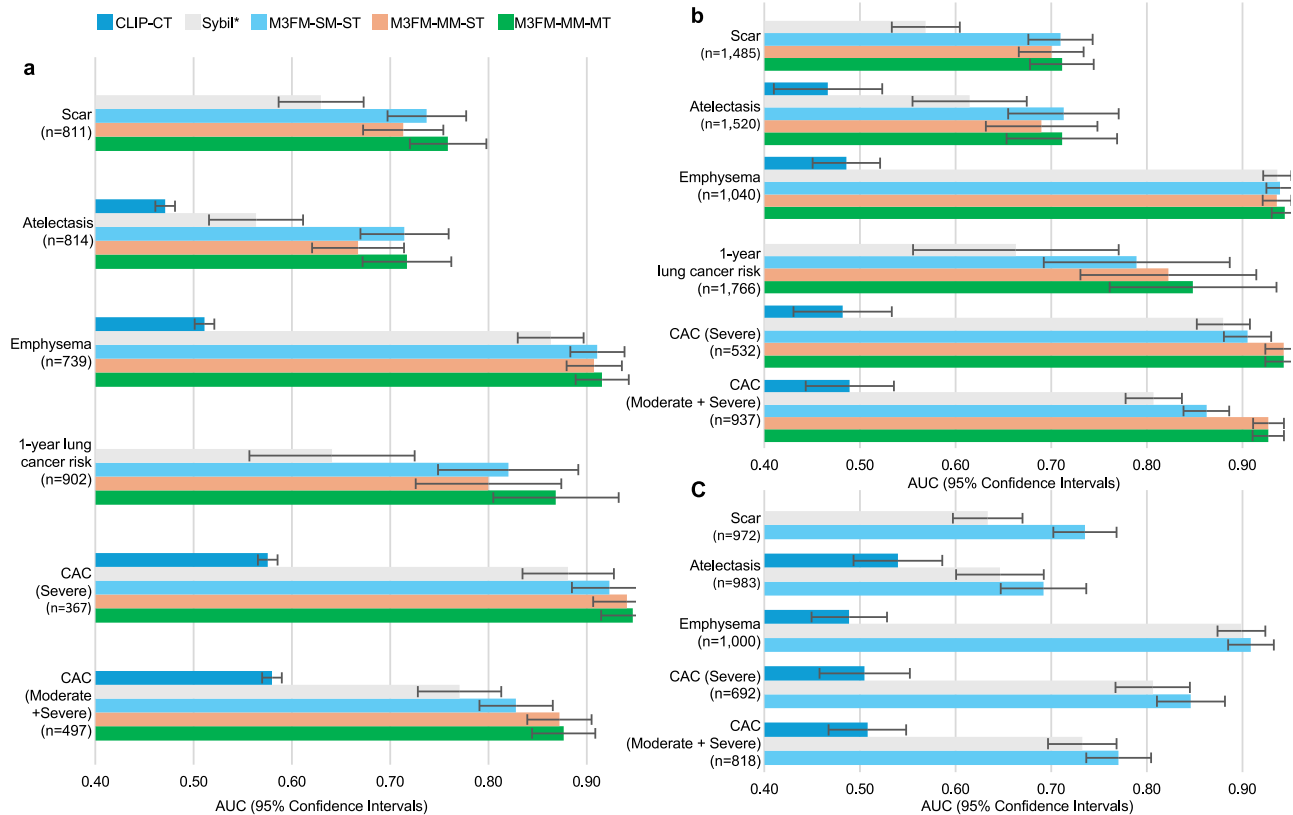
**Fig. 5 | Inspection of imaging and clinical data elements. a** Evaluation of voxel size embedding in computed tomography (CT) imaging. The Area Under the Curve (AUC) values and 95% confidence intervals for Medical Multimodal Multitask Foundation Model (M3FM) models are reported with and without embedding CT voxel sizes across various tasks. The AUC value and two-sided 95% confidence interval for each task were calculated from its entire test dataset. The error bars

indicate the two-sided 95% confidence intervals. Source data are provided as a Source Data file. **b** The attention maps of the task encoder for two cardiovascular disease (CVD) diagnosis examples, where the two cases were reported with significant CVD abnormalities. **c** The attention maps of the task encoder for two lung cancer risk prediction examples, where the pathology test results confirmed the lung cancer within one year following their low-dose CT lung cancer screenings.

single-task model (M3FM-MM-ST). Relative to M3FM-SM-ST, the M3FM-MM-ST model improved the AUC by 4.39% and 1.75% on the two CVD MGH datasets respectively. For the 1-year lung cancer risk prediction on the MGH dataset, the M3FM-MM-MT model improved the AUC by 20.80% against the previous model under the same experimental setting without using any bounding box annotations, improved the AUC by 4.85% over M3FM-SM-ST, and improved the AUC by 6.89%

over M3FM-MM-ST. On the MGH emphysema, atelectasis, and reticular opacities/honeycombing/fibrosis/scar datasets, the M3FM improved the AUC by 5.23%, 14.34%, and 12.91% relative to the previous model, and also achieved AUC improvements by 0.24% - 4.96% over M3FM-SM-ST and M3FM-MM-ST. For the CVD tasks on the MGH datasets, M3FM-MM-MT improved the AUC by 12% and 6.29% against the previous model, improved the AUC by 6.46% and 3.77% relative to M3FM-





**Fig. 6 | Evaluation of the Medical Multimodal Multitask Foundation Model (M3FM) and competing models on independent datasets.** Evaluation results of the M3FM variants including single-modality (SM), multimodality (MM), single task (ST), and multitask (MT), and competing models on the (a) MGH, (b) WFUSM datasets, and (c) WFUSM full-dose CT datasets in terms of Area Under the Curve

(AUC) and 95% confidence intervals. The AUC value and two-sided 95% confidence interval for each task were calculated from its entire test dataset. The error bars indicate the two-sided 95% confidence intervals. CAC Coronary Artery Calcification, which is a type of cardiovascular disease. Source data are provided as a Source Data file.

SM-ST, and had the essentially same results as M3FM-MM-ST. For the 1-year lung cancer risk prediction on the MGH dataset, the M3FM-MM-MT model improved the AUC by 18.54% relative to the previous model under the same experimental setting without using any bounding box annotations, improved the AUC by 5.91% over M3FM-SM-ST, and improved the AUC by 2.57% over M3FM-MM-ST; and M3FM-MM-ST improved the AUC by 3.24% over M3FM-SM-ST. On the WFUSM emphysema, atelectasis, and reticular opacities/honeycombing/fibrosis/scar datasets, the M3FM-MM-MT model improved the AUC by 0.78%, 9.65%, and 14.24% against the previous model. We further evaluated the generalizability of M3FM on the full-dose CT scans in Fig. 6c. It is observed that M3FM (M3FM-SM-ST) models trained with LDCT scans performed similarly on the diagnosis tasks of scar, atelectasis, and emphysema abnormalities, but had an evident performance drop for CVD diagnosis on full-dose CT scans. M3FMs still outperformed the competing models by 1% ~ 10% on all the compared tasks.

### M3FM enhancing out-of-distribution multimodal analysis

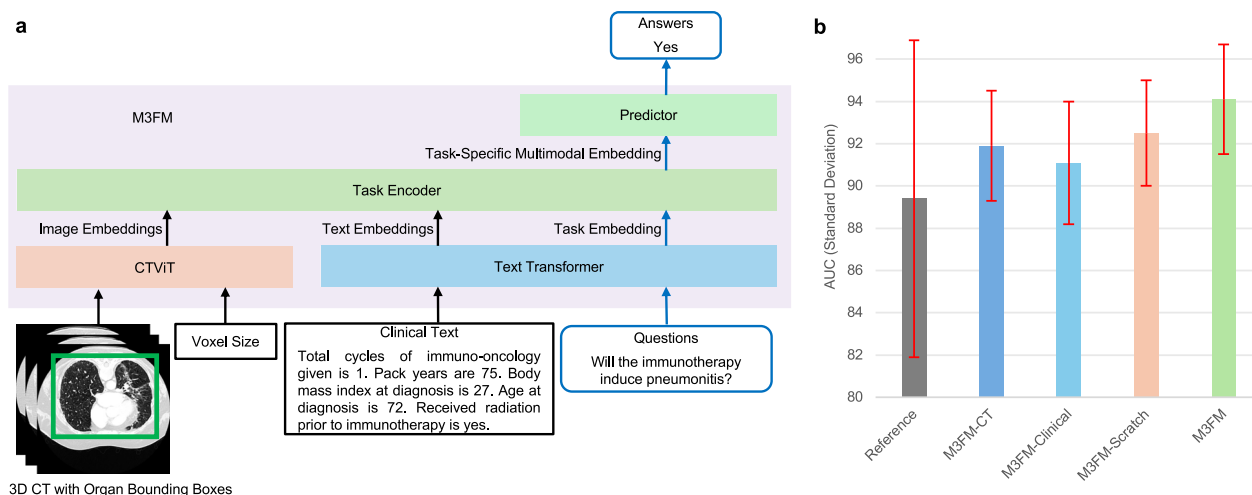
We further evaluated if M3FM, as a foundation model, facilitates out-of-distribution multimodal modeling as shown in Fig. 7. For this purpose, we fine-tuned M3FM to predict immunotherapy-induced pneumonitis from volumetric CT prior and the selected clinical data related to immunotherapy as described in the Results section. We used the method developed in WFUSM as the reference method<sup>51</sup> and compared different fine-tuned variants of the M3FM in terms of the average AUC and its standard deviation in five-fold cross-validation. The reference model used in WFUSM had  $0.894 \pm 0.075$  AUC by merging all radiomic and clinical features. Specifically, the reference model was

based on a nomogram to predict immunotherapy outcomes using features extracted from radiomic algorithms, a pre-trained ViT-base model, and clinical records. After feature selection, 20 radiomic features, 20 deep features, and 17 clinical features were used for the nomogram. The best result of our fine-tuned M3FMs was  $0.941 \pm 0.026$  of AUC, which achieved a 4.7% improvement over the competing model. The M3FM-CT model using CT data only had an AUC of  $0.919 \pm 0.026$ . The M3FM-Clinical model using clinical text only had a  $0.911 \pm 0.029$  AUC. The M3FM-Scratch without pretraining achieved  $0.925 \pm 0.025$  of AUC, in a favorable comparison with the competing model.

## Discussion

The contributions of the proposed M3FM can be summarized in two main aspects.

First, as the multimodal multitask foundation model for LCS, M3FM effectively encodes multimodal medical data including arbitrary combinations of multi-scale 3D tomographic images and various other clinical data, and flexibly performs multiple tasks via free-text prompting. In particular, our CT Vision Transformer (CTViT) is a unique component designed to perceive 3D CT images. CTViT can flexibly process multiple image sizes through our multi-scale linear tokenizer and disentangled physical size embedding mechanism. Our proposed self-supervised learning training algorithm facilitates the pre-training of the multi-scale CTViT on a large 3D CT dataset. To make M3FM scalable across multiple tasks, our distributed task-parallel training strategy assigns a single task to each device while allowing different devices to process different inputs/outputs for multitask parallel optimization.



**Fig. 7 | Transfer learning with the Medical Multimodal Multitask Foundation Model (M3FM).** **a** The same M3FM architecture was fine-tuned to perform the out-of-distribution immunotherapy prognosis task with three-dimensional (3D) computed tomography (CT) and clinical inputs. **b** Results on immunotherapy-induced pneumonitis using different methods, including the reference method, M3FM-CT which only takes CT as inputs, M3FM-Clinical which only takes the clinical data as

inputs, M3FM-Scratch that was trained from scratch without utilizing the pre-trained model, and M3FM that takes both CT and clinical data as inputs and was finetuned from the pre-trained model. The error bars represent mean AUC  $\pm$  standard deviations from five-fold cross-validation ( $n = 5$ ), with each fold for a distinct train/test split of the same dataset. CTViT Computed Tomography Vision Transformer. Source data are provided as a Source Data file.

Second, the whole workflow for the M3FM development is built for the clinically challenging scenario, from LCS multitask definition to multimodal data curation, from radiologists' interpreting procedure to the unified MQA framework, and from self-supervised pre-training to synergistic multitasking with high-dimensional multimodal data. In particular, our MQA framework is akin to how radiologists perform multiple tasks while considering multimodal data, naturally facilitating unified training and interactive inference. Importantly, the whole pipeline is designed with scalability, allowing M3FM to be readily scaled up by integrating more training datasets and undertaking a broader range of clinical tasks.

The M3FMs significantly outperforms the previous models developed on either single task and/or single modality, the CLIP-CT foundation model, and the generalist AI model on our curated large-scale OpenM3Chest datasets collected from multiple medical institutes. Our experimental results indicate that the larger M3FM produces better LCS outcomes in multimodal multitask settings. These positive outcomes underscore the importance of systematically collecting and curating large-scale, multimodal, multitask datasets for superior performance in LCS-related tasks.

The M3FMs effectively encode various combinations of multimodal inputs. The results in Table 1 show that the multimodal data are particularly helpful for improving CVD diagnosis and CVD mortality risk prediction with M3FM. However, some other tasks, such as lung nodule detection and characterization did not benefit from additional clinical data types, suggesting that the imaging data is enough for these tasks. On the other hand, the evaluation results in Table 2 and Supplementary Tables 4 and 5 show that training M3FM with additional clinical input data types would not degrade the performance in comparison with the single-modality LDCT models on the tasks that were not benefited from the clinical priors. In other words, there are no constraining effects of multimodal learning on the representations of the single-mortality models trained with LDCT scans only. Therefore, incorporating multimodal data as inputs is generally beneficial or at least does not harm model performance. In this context, M3FM also provides a flexible framework to study the correlation between the model performance and different combinations of multimodal inputs on each specific task for optimized selection of multimodal datasets.

The M3FMs flexibly synergize different medical tasks. Theoretical studies<sup>52,53</sup> have indicated why learning multiple tasks jointly is beneficial over learning each task in isolation through analysis of the upper error bound conditioned on the number of multiple tasks and the average number of data points per task. In the real-world LCS applications, our results have shown that multitask learning generally achieved better generalizability on both the NLST test datasets and independently collected evaluation datasets, which are consistent with the theoretical findings. Interestingly, the empirical results further suggest that the LCS tasks benefited from multitask learning have more imbalanced labels. This seems heuristic, since when labels are sparse for a specific task, more synergy should be leveraged by learning from other related tasks. In other words, multitask learning promises to alleviate the over-fitting and improve the generalizability via a multi-task-based regularization effect especially when a task of interest has a limited number of samples<sup>52,53</sup>. Given that many clinical tasks involve highly imbalanced datasets with few labels of positive or certain type disease, our results mean that multitask learning is favorable in building foundation models in real-world scenarios. Overall, the results in Table 1 demonstrate that M3FM optimized with multitask learning outperformed the corresponding models that were separately optimized for individual tasks. These results indicate that for intrinsically-related tasks if their multimodal data and clinical labels are simultaneously collected, such as in the NSLT trial for LCS, M3FM can synergistically integrate multiple tasks that take various scales of imaging data and different multimodal inputs for improved performance using our proposed distributed task-parallel training approach. However, when non-trivial efforts are required to collect additional multitask datasets for a particular task at hand, there would be a trade-off between performance improvement (especially if the improvement is marginal) and the associated cost of constructing additional datasets. In the latter case, M3FM provides a useful framework to study this trade-off by performing ablation studies as shown in Table 1 on relatively small pilot datasets prior to collecting large-scale datasets.

Nevertheless, there are certain limitations to current M3FM results. The above evaluation was retrospective and offline rather than in a prospective, real-world reporting and patient management environment. We did not test the clinical impact of M3FM either in radiology or post-radiology care scenarios. Likewise, we did not evaluate the most

effective method of information display to improve decision-making with multimodal information without inundating physicians and compromising their workflow efficiency. While our experimental results affirm that larger-scale models yield better multimodal multitask performance, the performance gain when upgrading from M3FM-Large to M3FM-Huge is less impressive than when upgrading from M3FM-Base to M3FM-Large. We believe that this limited improvement could be substantially attributed to the size and quality of the current datasets, and with even larger and better datasets we expect to have a higher performance gain, by the scaling laws. Although the current M3FM models can perform a major set of LCS-related tasks covering all radiological labels in NLST, it is limited to predicting the abnormalities within the lungs and heart since the available labels are mainly targeted to these regions in our collected datasets. However, thanks to the built-in capability of multi-scale deep CT image analysis, it is straightforward to extend our M3FM models to handle other opportunistic findings<sup>54–57</sup> by focusing on the involved regions coupled with relevant clinical data and corresponding labels. It is also observed that the AI model performance faces significant challenges, especially for real-world clinical tasks with highly imbalanced data distributions and small abnormality regions in the 3D input that contains no location annotations. By addressing these limitations, there are opportunities for clinical impacts of M3FM. After its regulatory clearance and integration into real-time clinical workflows, M3FM can provide a dynamic, and customizable dashboard for information summary and decision-making. For example, during the reporting of lung findings, it would be helpful to have a display of M3FM-derived results, previously reported lung-specific clinical findings (such as Chronic Obstructive Pulmonary Disease (COPD), prior lung nodules, and LungRADS categories); and during cardiovascular field reporting, to have a display of M3FM-based estimates and cardiac risk factors from the past medical history.

As discussed in<sup>58</sup>, there are specific concerns for AI in medicine, such as generalizability, explainability, adaptability, etc. This study has demonstrated initial efforts in addressing such AI-specific concerns.

The M3FM models show better generalizability to independent WFUSM and MGH datasets than the SOTA models. It is worth mentioning that the independent evaluation is prospective in terms of the data collection date. Our experimental results show that M3FMs achieved consistently and significantly better results by up to 20+% than the previous models trained in the same setting. However, for some tasks, multi-modality modeling could decrease the generalizability relative to single-modality modeling. This might be due to the variability in data collection procedures and standards. Thus, it is important to design a standardized and robust data resourcing and collecting pipeline. In all our experiments, multitask learning consistently improved generalizability.

The M3FMs are capable of identifying informative clinical elements both quantitatively and qualitatively, which offers a high-level explainability. It is achieved with the MQA framework and the attention mechanism. Specifically, the MQA framework naturally allows users to examine the response changes to different combinations of imaging and clinical data, and thus the informative clinical elements can be identified as those contributing to statistically high prediction accuracy. Our M3FMs have uncovered a strong positive correlation between CVD diagnosis, CVD mortality prediction and the historical presence of heart disease/attacks, hypertension, stroke, and diabetes through ablation inference. Assuming this discovery is not a piece of common knowledge, M3FM could contribute to important biomedical insights. Quantitatively, attention maps can be visualized for both image and text inputs through the attention mechanism, illuminating the elements that correlate with predictions. This visualization offers a certain interpretability of M3FM in terms of the relationships between clinical data and diseases.

The M3FM has the adaptability to significantly improve multimodal modeling for out-of-distribution tasks through transfer

learning. A key feature of foundation models is their ability to aid tasks beyond those defined by the training datasets. In this study, we fine-tuned our M3FM for immunotherapy prognosis prediction, an out-of-distribution task characterized by entirely different clinical inputs. Our experiments demonstrated that the pre-trained M3FM model can handle the out-of-distribution task with a high performance on a relatively small dataset. This capability is particularly valuable when expanding some clinical datasets is challenging due to data rarity and associated costs.

In conclusion, the unified architecture and exceptional performance of the M3FMs herald a promising avenue for leveraging multimodal data to perform multitasks in developing AI-empowered, specialty-oriented superior healthcare solutions. Within the scope of LCS in particular, we have demonstrated the feasibility of translating the M3FM model on our collaborative clinical sites, broaden and refine LCS implementation, and ultimately reduce lung cancer mortality. Hopefully, our M3FM system would become an effective platform to accommodate more medical tasks with diverse multimodal data combinations, from specialized to increasingly more generalized medical AI models.

## Methods

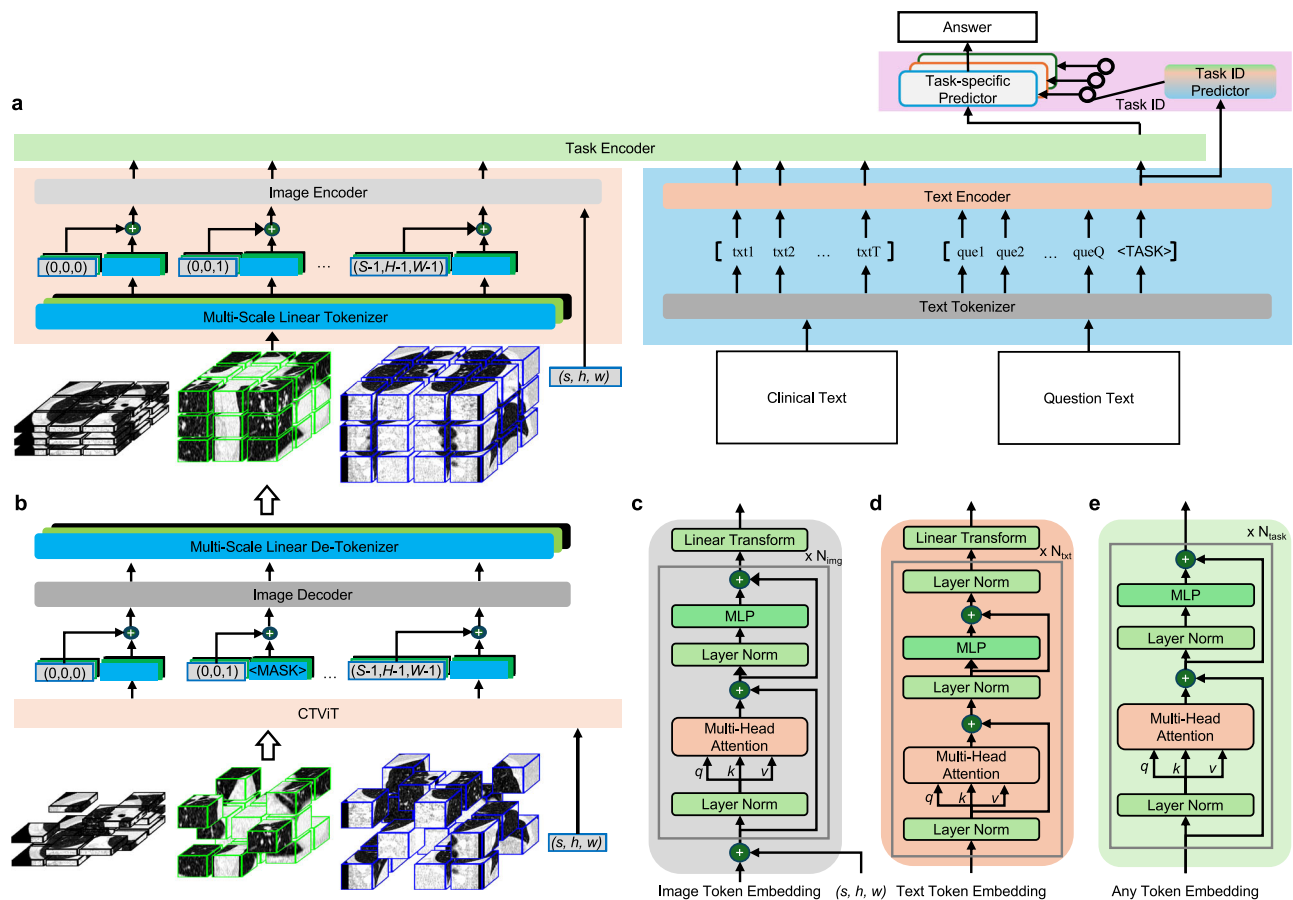
### Ethics statement

This study was conducted in accordance with all relevant ethical regulations after the Institutional Review Board (IRB) approvals of WFUSM (IRB approval number: IRB00002960) and MGH (IRB approval number: 2020P003950). Both IRBs granted a waiver of informed consent as all data in this study were retrospective and de-identified, which is in accordance with 45 CFR 46.116(f). Access to and use of the MIDRC and NLST datasets were conducted in compliance with their respective data use agreements and adhered to applicable data privacy standards.

### Medical Multimodal-Multitask Dataset Construction

Figure 2a presents our general workflow for constructing the multimodal multitask medical datasets, which consists of the four main steps: (1) medical task definition; (2) task-specific multimodal data collection; (3) multimodal data processing and alignment; and (4) MQA construction. The details for the first two steps are described in the Results section. Here we describe the third and fourth steps.

Multimodal data processing is to select qualified multimodal data and prepare them for the alignment with each clinical task of interest, including CT data processing, clinical data processing, and ground-truth labeling. In CT data processing, we localize the sub-volumes that mainly contain the task-relevant regions in the 3D CT volume using a segmentation model<sup>59</sup>. Specifically, we segment three parts, i.e., left lung, right lung, and heart regions consisting of the myocardium, left/right atrium, left/right ventricle, and pulmonary artery. It is worth underlining that the precision of the segmentation results does not need to be high. Our primary objective is to utilize rectangular boxes to wrap the segmented areas, ensuring that task-relevant sub-volumes are included and extraneous regions are disregarded. We excluded the CT series having less than 64 axial slices in all collected datasets. For each CT series, the reconstruction voxel sizes are recorded in the axial, coronal, and sagittal dimensions, and used as input to the CTViT. The clinical data processing is to represent various combinations of clinical data as free text. We have established a specific sentence format for each clinical element, as detailed in Supplementary Table 6. The final free-text clinical data for each examination is constructed by aggregating the sentences corresponding to all available and labeled clinical data, as illustrated by the text inputs in Figs. 1 and 7. For ground-truth labeling, we first extract task-specific labels from different sources (see Fig. 3a for specific label sources of each task) and then combine all information as the label. The details for each task ground-truth label calculation are described in Supplementary Table 7. Next, we align the clinical data in free text, the CT data with segmented parts and physical size, and the labels in all exams for each task. In particular, each task



**Fig. 8 | M3FM architecture.** **a** The overall M3FM architecture with Computed Tomography Vision Transformer (CTViT), text transformers, encoders, and predictors. **b** CTViT pretraining. **c** image encoder. **d** text encoder. **e** task encoder.  $S, H, W$  denote the sizes of CT volume in pixels.  $s, h, w$  denote the sizes in millimeters of a voxel in CT.  $txt1, txt2, \dots, txtT$  denote the clinical text tokens.  $que1, que2, \dots, queQ$  denote the question text tokens.  $\langle MASK \rangle$  and  $\langle TASK \rangle$  are the special text

tokens. Task ID is the identifier of a specific task.  $q, k, v$  denote the queries, keys, and values in the Transformer model.  $N_{img}, N_{txt}, N_{task}$  denote the number of Transformer blocks in the Image Encoder, Text Encoder, and Task Encoder, respectively. Different colors with Multi-Scale Linear Tokenizers/De-Tokenizers match different scales.

anatomically corresponds to its segmented CT sub-volume. The key consideration is to remove irrelevant image regions to reduce the computational cost while keeping the original information. The task-specific CT sub-volumes are illustrated in the image input column of Fig. 3a. Specifically, the left or right lung sub-volume is used as the image input for lung nodule detection and characterization. We fixed the number of slices to 16 for each input considering that a lung nodule is usually tiny relative to the whole lungs. For the nodule-presented case, the location labels of the left/right lung, slice number, and the bounding box coordinates are used to crop the target sub-volume. For the non-presented nodule case, the input sub-volume is randomly cropped within the segmented lung regions. The CVD tasks use a 3D box wrapping the heart as the image input. For lung cancer risk prediction, we separately input the left and right lungs. In the training datasets, the location of the left and/or right lung where lung cancer is presented is required to align each lung with the risk labels. In the inference stage, the location of lung cancer is not required because the ground-truth labels are at the patient level. The predicted lung cancer risks are the maximum of the scores of two lungs for each patient. For all other chest diagnosis tasks, a 3D rectangle box wrapping both lungs is used as the image input. Subsequently, the MQA construction is to create questions and answers for each specific task with the aligned multimodal data, and the resulting MQA datasets define the model's input and output formats. In Fig. 3a, one example question and the corresponding answer candidates are presented for each task. In the

training stage, ten different questions for each task were used as shown in Supplementary Table 8. Note that there is no need for radiologists to make labeling efforts across the whole workflow so that large-scale medical multimodal multitask datasets can be cost-effectively constructed.

### M3FM

**Overall architecture.** Our medical multimodal multitask foundation model is designed to effectively encode multimodal data and flexibly perform multitasks via text prompting in a unified and scalable fashion. As shown in Fig. 1a, M3FM consists of four main components: CTViT, text Transformer, task encoder, and predictors. The key details of each component are given in Fig. 8. CTViT takes volumetric CT images of varying sizes as inputs, extracts multi-scale image patches, and computes discriminative features of these patches. The text Transformer produces the embedding of clinical text and the embedding of textual questions respectively. Given any combination of image, text, and task token embeddings, the task encoder extracts the task-specific features corresponding to the special  $\langle TASK \rangle$  token. Finally, the task-specific predictor outputs the final answer from the task-specific features of the integrated multimodal data. In the following, we will describe each component in detail.

**CTViT.** CTViT extracts embedding features of multi-scale 3D CT volumes with size awareness. CTViT has two parts: a multi-scale CT



tokenizer and an image encoder. **To process a 3D CT scan, we divide each image volume into non-overlapped 3D patches as in<sup>60</sup>. Each 3D patch is referred to as an image token.** Since various diseases are at different scales in the CT images, we design a multi-scale CT tokenizer, which consists of multiple linear embedding layers corresponding to different sizes of image patches, as shown in Fig. 8a. Each embedding layer has a linear transformation and a set of learnable positional embeddings. Each image token embedding is the sum of its linear transformation and the positional embedding. All sizes of image tokens are mapped to the same image embedding space. Inspired by<sup>61</sup>, we decompose the 3D position embedding into two parts indexing in-plane and through-plane positions respectively. In other words, we have two positional embeddings: one for the 2D space within each slice and the other for the 1D range of slice position. The 3D positional embedding is the sum of them. By doing so, the number of learned parameters can be reduced. Figure 8c shows the image encoder in detail. Different CT scans may have different physical sizes specific to the individual patient size. The physical size is an important factor in some clinical tasks. Thus, we encode the physical size with sine-cosine functions of different frequencies and add it to the image token embedding. Then, the image encoder was implemented as in the plain ViT<sup>39</sup> that consists of multiple self-attention Transformer layers and a subsequent linear transformation layer that maps the image embedding space to the task embedding space. By disentangling physical size from the image content, we can flexibly perceive any size of CT volumes with size awareness without resampling CT volumes to have a consistent image tensor across different inputs.

Empirically, we predefined four scales of embedding layers; i.e., the volume size of  $16 \times 448 \times 320$  with the patch size of  $4 \times 16 \times 16$ , the volume size of  $128 \times 448 \times 320$  with the patch size of  $16 \times 16 \times 16$ , the volume size of  $128 \times 192 \times 224$  with the patch size of  $16 \times 16 \times 16$ , and the volume size of  $128 \times 320 \times 448$  with the patch size of  $16 \times 16 \times 16$ , to encode lung nodule, heart, lung cancer, and other chest abnormalities respectively. It is worth noting that any prior attention to sub-volumes can be further applied within each scale by adding the attention masks to all self-attention layers as what is done in NLP models<sup>62</sup>. We used the bounding boxes of lungs to make the model attend to lungs only in predicting and characterizing lung nodules. For M3FM-Base, M3FM-Large, and M3FM-Huge, the numbers of transformer layers are 12, 24, and 32, and the sizes of image token embeddings are 768, 1,024, and 1,280, respectively.

**Text transformer.** Any decent language model can be utilized as the text Transformer in M3FM. Here the text Transformer was implemented as a Byte-level Byte-Pair-Encoding (BBPE) tokenizer<sup>62</sup>, a text encoder consisting of the original Transformer layers<sup>63</sup> and a linear transformation layer, as shown in Fig. 8d. On one hand, the text encoder encodes patient-specific clinical information, such as demographics, smoking history, disease history, cancer history, and other clinical data, which are free text; for example: The patient is 56.0 years old. Gender is Female. Ethnicity is neither Hispanic nor Latino. Height is 60.0 inches. Weight is 105.0 pounds. Education is associate degree/some college. Former smoker. Smoking duration is 38.0 pack years. Smoking intensity is 20.0 cigarettes per day. 2.0 years since quit smoking. The patient had asthma (childhood) diagnosed at 7.0 years old. The patient had hypertension diagnosed at 53.0 years old. The patient had pneumonia diagnosed at 50.0 years old. Patient's brother(s) (including half-brothers) have lung cancer. On the other hand, the text encoder encodes free-text task instructions/questions, which are used as the input of the task encoder to extract task-specific embedding features from the multimodal data; for example: Is there any significant cardiovascular abnormality? and Predict the lung cancer risk over 6 years. This approach allows for embedding any combination of clinical information through free-text prompting, regardless of order. The control signals for specific tasks are then extracted from the

text prompts by the same text encoder. Again, the linear transformation maps the text embedding space to the task embedding space. For all our M3FMs, the number of Transformer layers is 12, and the size of text token embeddings is 768. Additionally, we comprehensively compared different methods for encoding clinical data in the format of an array, format-fixed text, and free-form text respectively. The results show that encoding clinical data with either format-fixed or free-form text achieved better results than that with array data. Also, encoding clinical data into format-fixed and free-form text achieved similar results. See Supplementary Methods and Supplementary Table 9 for details.

**Task Encoder.** Figure 8e illustrates the task encoder to extract task-specific embedding features from the multimodal token embeddings, given the special `< TASK >` token embedding. The task encoder was implemented with multiple Transformer layers, treating all tokens as a single input sequence. Note that only the special `< TASK >` token is forwarded to the task encoder and the rest of the question tokens are ignored, as we empirically found that the other tokens did not increase the performance in practice. The special token embedding from the final Transformer layer serves as the task-specific embedding feature that integrates all multimodal data. For M3FM-Base, M3FM-Large, and M3FM-Huge, the number of Transformer layers is 4 in every case, and the sizes of task token embeddings are 768, 1024, and 1280, respectively.

### Predictors

The Predictors map task-specific embedding features to answers. In this study, we found that the task-specific predictor can be automatically selected well through the Task ID Predictor, which takes the `< TASK >` embedding corresponding to the question text. We implemented all Predictors including the Task ID Predictor as a two-layer MLP. Different tasks may have different Predictors or shared Predictors for the same output dimension, such as Yes or No answers. Similar to language models that regard text generation as the token classification over a vocabulary, we formulate our answer prediction as a classification problem over the predefined answer candidates, as summarized in Fig. 3a, except for a six-year lung cancer risk prediction. Similar to<sup>14</sup>, we formulate lung cancer risk prediction as a hazard regression problem. It is worth mentioning that more types of prediction tasks, such as image segmentation and object detection, can be incorporated into M3FM by adding the corresponding lightweight task-specific predictors as demonstrated in our previous study<sup>64</sup>.

### Self-supervised pretraining

A key step to optimize large models is self-supervised pretraining with large unlabeled datasets. In this study, we adapted the masked auto-encoder method<sup>61,65</sup> to pretrain our CTViT on our OpenM3Chest pretraining dataset. Figure 8b shows the CTViT pretraining architecture, which consists of CTViT, an image decoder, and a multi-scale linear de-tokenizer. The image encoder was optimized by predicting masked cubes (85%) from a small number of visible cubes (15%). To reduce the memory overhead, only some selected slices along the longitude direction were predicted while recovering each 3D patch. We pre-trained CTViT with the pre-defined multi-scale 3D CT volumes and a set of data augmentation operations, including random cropping, rotation, resizing, and perturbed display windowing. The text Transformer in M3FM was initialized with the off-the-shelf RoBERTa model pre-trained via masked language modeling and then trained end-to-end<sup>62,66</sup>.

### Multitask learning

After self-supervised pretraining, M3FM can be trained with any combination of different tasks with properly selected multimodal datasets by optimizing multitask loss functions simultaneously. We

used the sigmoid cross-entropy loss function for the CVD mortality risk and lung cancer risk prediction tasks and the softmax cross-entropy loss for all other tasks. As the number of tasks increases, there is a significant rise in computational cost. To address this problem, we designed a distributed task-parallel (DTP) training strategy. TDP assigns each computing device with a single task and a single data loader while the total number of training samples remains fixed across all devices for each task. Since M3FM is a unified model capable of handling various tasks, despite differences in input and output dimensions, gradients computed across all tasks can be readily accumulated, enabling simultaneous parameter optimization.

### Transfer learning

M3FM is designed for adaptability and generalization, enabling the enhancement of out-of-distribution task performance through transfer learning. This capability extends to out-of-distribution tasks with varying image input dimensions, clinical data types, and output dimensions. To accommodate different image dimensions, the addition of a linear embedding layer suffices. For diverse clinical datasets, we can simply describe involved clinical data in free text to the model, without needing any modification on the M3FM architecture, as shown in Fig. 7a. Specifically, adjusting to different output dimensions requires only the inclusion of a lightweight predictor. Consequently, M3FM can be easily fine-tuned to enable out-of-distribution tasks by leveraging the pre-trained model parameters.

### Training details

We used the AdamW optimizer<sup>67</sup>, cosine decay learning rate schedule<sup>68</sup>, weight decay of 0.05, and automatic mixed precision in PyTorch for training all models. In pretraining CTViT, the CT volume was randomly scaled by the factor of [0.5, 2], [0.6, 1.4], and [0.6, 1.4] in axial, coronal, and sagittal dimensions, with the voxel size accordingly calculated. Then, on each GPU a single input size was randomly chosen from a predefined set of input sizes. The input region of the chosen size on the current scale was randomly cropped within the whole CT volume, not limited to the lung and heart regions. The CTViT was pre-trained for 200K iterations with 10K warmup iterations, the decoder depth was 2, the voxel values were normalized within each cube in calculating the MSE loss, the batch size was 192 and the learning rate was  $3.75 \times 10^{-4}$ . In training task-specific models including the transfer learning, the batch size was 12, and the number of training iterations was 30K with 2K warm-up iterations. The learning rate was  $2 \times 10^{-4}$ , and the layer-wise learning rate decay of 0.95 was used. In multitask training, the total batch size was 972, including 12 samples for each of the 17 tasks and 768 samples for the clinical information retrieval tasks. All CT inputs had a random HU range perturbation, random rotation degrees, and random padding in training the M3FM models, with the corresponding hyperparameters in training for different tasks described in Supplementary Table 10. Each clinical data element was randomly included with a probability of 0.8.

### Hardware requirement

All our models were trained on the AiMOS Supercomputer in the Center for Computational Innovation at Rensselaer Polytechnic Institute ([https://docs.cci.rpi.edu/clusters/DCS\\_Supercomputer/](https://docs.cci.rpi.edu/clusters/DCS_Supercomputer/)). For CTViT pretraining and multi-task training, we used 192 NVIDIA Tesla V100 GPUs with 32 GiB of memory each, i.e., 6 GPUs per node  $\times$  32 nodes. The CTViT pretraining took around 60 hours. The multi-task training took about 30 hours. For all single-task training and finetuning, we used 12 NVIDIA Tesla V100 GPUs with 32 GiB of memory each, i.e., 6 GPUs per node  $\times$  2 nodes. The single-task training took about 22 hours.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The OpenM3Chest data generated in this study have been deposited in the Zenodo database under the accession code [14363994](https://doi.org/10.5281/zenodo.14363994). The corresponding image data in the OpenM3Chest can be obtained from NLST and MIDRC. The deidentified datasets from WFUSM and MGH only allow restricted access according to the requirements of the institutional review board-approvals and the data sharing regulations as WFUSM and MGH forbid open access to their patients' data. Access can be obtained after the IRB and Data Sharing Committee approvals at the WFUSM, MGH and the requesting institution (details on how to request access are available from Dr Christopher Whitlow at WFUSM and Dr Mannudeep Kalra at MGH). Source data are provided with this paper.

### Code availability

The code used in this study is publicly available at <https://github.com/niuchuangnn/M3FM> under the MIT License<sup>69</sup>. The license text is provided in the repository's LICENSE file. Attribution and usage terms comply with the licensing agreement.

### References

1. American Lung Association: Lung Cancer Trends Brief. <https://www.lung.org/research/trends-in-lung-disease/lung-cancer-trends-brief>. Accessed: 2024-12-11 (2024).
2. National Lung Screening Trial Research Team Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Eng. J. Med.* **365**, 395–409 (2011).
3. Koning, H. J. et al. Reduced lung-cancer mortality with volume ct screening in a randomized trial. *N. Engl. J. Med.* **382**, 503–513 (2020).
4. Fedewa, S. A. et al. State variation in low-dose computed tomography scanning for lung cancer screening in the united states. *JNCI: J. Natl Cancer Inst.* **113**, 1044–1052 (2021).
5. Jonas, D. E. et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the us preventive services task force. *JAMA* **325**, 971–987 (2021).
6. Rivera, G.A., Wakelee, H., In: Ahmad, A., Gadgeel, S. (eds.) Lung Cancer in Never Smokers, pp. 43–57. Springer, Cham (2016).
7. Triplette, M. et al. Patient identification of lung cancer screening follow-up recommendations and the association with adherence. *Ann. Am. Thorac. Soc.* **19**, 799–806 (2022).
8. Lin, Y. et al. Patient adherence to lung ct screening reporting & data system-recommended screening intervals in the united states: A systematic review and meta-analysis. *J. Thorac. Oncol.* **17**, 38–55 (2022).
9. Núñez, E. R. et al. Adherence to follow-up testing recommendations in us veterans screened for lung cancer, 2015–2019. *JAMA Netw. Open* **4**, 2116233–2116233 (2021).
10. Glover IV, M. et al. Socioeconomic and demographic predictors of missed opportunities to provide advanced imaging services. *J. Am. Coll. Radiol.* **14**, 1403–1411 (2017).
11. Tseng, C.-H. et al. The relationship between air pollution and lung cancer in nonsmokers in taiwan. *J. Thorac. Oncol.* **14**, 784–792 (2019).
12. Wang, G. X. et al. Barriers to lung cancer screening engagement from the patient and provider perspective. *Radiology* **290**, 278–287 (2019).
13. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
14. Mikhael, P. G. et al. Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *J. Clin. Oncol.* **41**, 2191–2200 (2023).
15. Ruparel, M. et al. Evaluation of cardiovascular risk in a lung cancer screening cohort. *Thorax* **74**, 1140–1146 (2019).

16. Chao, H. et al. Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. *Nat. Commun.* **12**, 2963 (2021).
17. Bommasani, R. et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).
18. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901 (2020).
19. Anil, R. et al. Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023).
20. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
21. Dehghani, M. et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512 (2023).
22. Kirillov, A. et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026 (2023).
23. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763 (2021).
24. Chen, X. et al. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations* (2023).
25. Girdhar, R. et al. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190 (2023).
26. Li, Y. et al. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* **15**, 1–12 (2023).
27. Han, T. et al. MedAlpaca—An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv preprint arXiv:2304.08247 (2023).
28. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. PMC-LLaMA: toward building open-source language models for medicine. *J. Am. Med. Inform. Assoc.* **31**, 1833–1843 (2023).
29. Toma, A. et al. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. arXiv preprint arXiv:2305.12031 (2023).
30. Xiong, H. et al. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097 (2023).
31. Wang, H. et al. Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023).
32. Ichinose, A. et al. Visual grounding of whole radiology reports for 3d ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 611–621 (2023).
33. Bannur, S. et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027 (2023).
34. Hamamci, I.E. et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. arXiv preprint arXiv:2403.17834 (2024).
35. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
36. Saab, K. et al. Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024).
37. Yang, L. et al. Advancing multimodal medical capabilities of gemini. arXiv preprint arXiv:2405.03162 (2024).
38. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
39. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
40. Chen, W. et al. Machine learning with multimodal data for covid-19. *Heliyon* (2023).
41. Medical Imaging and Data Resource Center Team: Medical Imaging and Data Resource Center (MIDRC). <https://www.midrc.org/>. Accessed: 2024-12-11 (2024).
42. Zeleznik, R. et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nat. Commun.* **12**, 715 (2021).
43. Guo, H., Kruger, U., Wang, G., Kalra, M. K. & Yan, P. Knowledge-based analysis for mortality prediction from ct images. *IEEE J. Biomed. health Inform.* **24**, 457–464 (2019).
44. Gunraj, H., Sabri, A., Koff, D. & Wong, A. COVID-Net CT-2: enhanced deep neural networks for detection of COVID-19 from chest ct images through bigger, more diverse learning. *Front. Med.* **8**, 729287 (2022).
45. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H. nnde-tecton: A self-configuring method for medical object detection. In *Medical Image Computing and Computer Assisted Intervention*, pp. 530–539 (2021).
46. OpenAI: Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed: 2024-12-11 (2024).
47. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* **143**, 29–36 (1982).
48. Kaplan, J. et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020).
49. Zhang, J. et al. Revisiting the trustworthiness of saliency methods in radiology AI. *Radiology: Artif. Intell.* **6**, 220221 (2023).
50. Darcet, T., Oquab, M., Mairal, J. & Bojanowski, P. Vision transformers need registers. In *International Conference on Learning Representations* (2024).
51. Lyu, Q., Yuan, H., Lin, Z., Ponnatapura, J. & Whitlow, C. T. A novel prediction model for immunotherapy induced pneumonitis prediction based on chest ct and electronic health record. *medRxiv* <https://doi.org/10.1101/2024.10.14.24315487> (2024).
52. Maurer, A., Pontil, M. & Romera-Paredes, B. The benefit of multitask representation learning. *J. Mach. Learn. Res.* **17**, 1–32 (2016).
53. Zhang, Y. & Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. data Eng.* **34**, 5586–5609 (2021).
54. Pickhardt, P. J. et al. Opportunistic screening: radiology scientific expert panel. *Radiology* **307**, 222044 (2023).
55. Cheng, X. et al. Opportunistic screening using low-dose ct and the prevalence of osteoporosis in china: a nationwide, multicenter study. *J. Bone Miner. Res.* **36**, 427–435 (2020).
56. Xu, K. et al. Ai body composition in lung cancer screening: added value beyond lung cancer detection. *Radiology* **308**, 222937 (2023).
57. Chen, X. et al. Elevated prevalence of moderate-to-severe hepatic steatosis in world trade center general responder cohort in a program of ct lung screening. *Clin. imaging* **60**, 237–243 (2020).
58. Wang, G. Making “CASES” for AI in medicine. *BME Front.* **5**, 0036 (2024).
59. Wasserthal, J. et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artif. Intell.* **5**, 230024 (2023).
60. Niu, C. & Wang, G. Unsupervised contrastive learning based transformer for lung nodule detection. *Phys. Med. Biol.* **67**, 204001 (2022).
61. Feichtenhofer, C. et al. Masked autoencoders as spatiotemporal learners. *Adv. neural Inf. Process. Syst.* **35**, 35946–35958 (2022).
62. Liu, Y. et al. Roberta: A robustly optimized Bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019).
63. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
64. Niu, C., Wang, G.: Ct multi-task learning with a large image-text (lit) model. arXiv preprint arXiv:2304.02649 (2023).
65. He, K. et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009 (2022).

66. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186 (2019).
67. Loshchilov, I., Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2018).
68. Loshchilov, I., Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations* (2016).
69. Niu, C. et al. Medical Multimodal Multitask Foundation Model for Lung Cancer Screening. Code available at M3FM repository <https://doi.org/10.5281/zenodo.14366852> (2024).

## Acknowledgements

This work was partly supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under awards R01EB031102 (G.W.) and R01EB032716 (G.W.). The imaging and associated clinical data downloaded from MIDRC (The Medical Imaging and Data Resource Center) and used for research in this study was made possible by NIBIB of NIH, under contract 75N92020D00021. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank the National Cancer Institute (NCI) for access to its data collected in the National Lung Screening Trial (NLST). The statements in this paper are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

## Author contributions

C.N. and G.W. initiated the project. G.W. supervised the project with M.K.K. and C.T.W. who were especially instrumental in clinical task definition, multimodal data analysis, clinical validation, and testing. C.N. curated the multimodal multitask datasets, designed and implemented the model architecture and training/inference algorithms, conducted all experiments, and produced results in figures and tables. Q.L. and J.T. collected multimodal datasets from WFUSM. Q.L. collected and processed the out-of-distribution dataset. P.K. collected and annotated multimodal datasets from MGH. C.D.C. provided technical support on high-performance computing. C.N., P.Y., M.K.K., C.T.W., and G.W. engaged in extensive discussions and wrote the manuscript. All the authors reviewed and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-56822-w>.

**Correspondence** and requests for materials should be addressed to Mannudeep K. Kalra, Christopher T. Whitlow or Ge Wang.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025