Pan-mediastinal neoplasm diagnosis via nationwide federated learning: a multicentre cohort study

Ruijie Tang, Hengrui Liang*, Yuchen Guo*, Zhigang Li*, Zhichao Liu*, Xu Lin*, Zeping Yan, Jun Liu, Xin Xu, Wenlong Shao, Shuben Li, Wenhua Liang, Wei Wang, Fei Cui, Huanghe He, Chao Yang, Long Jiang, Haixuan Wang, Huai Chen, Chenguang Guo, Haipeng Zhang, Zebin Gao, Yuwei He, Xiangru Chen, Lei Zhao, Hong Yu, Jian Hu, Jiangang Zhao, Bin Li, Ci Yin, Wenjie Mao, Wanli Lin, Yujie Xie, Jixian Liu, Xiaoqiang Li, Dingwang Wu, Qinghua Hou, Yongbing Chen, Donglai Chen, Yuhang Xue, Yi Liang, Wenfang Tang, Qi Wang, Encheng Li, Hongxu Liu, Guan Wang, Pingwen Yu, Chun Chen, Bin Zheng, Hao Chen, Zhe Zhang, Lunqing Wang, Ailin Wang, Zongqi Li, Junke Fu, Guangjian Zhang, Jia Zhang, Bohao Liu, Jian Zhao, Boyun Deng, Yongtao Han, Xuefeng Leng, Zhiyu Li, Man Zhang, Changling Liu, Tianhu Wang, Zhilin Luo, Chenglin Yang, Xiaotong Guo, Kai Ma, Lixu Wang, Wenjun Jiang, Xu Han, Qing Wang, Kun Qiao, Zhaohua Xia, Shuo Zheng, Chenyang Xu, Jidong Peng, Shilong Wu, Zhifeng Zhang, Haoda Huang, Dazhi Pang, Qiao Liu, Jinglong Li, Xueru Ding, Xiang Liu, Liucheng Zhong, Yutong Lu, Feng Xu, Qionghai Dai, Jianxing He*

Summary

Background Mediastinal neoplasms are typical thoracic diseases with increasing incidence in the general global population and can lead to poor prognosis. In clinical practice, the mediastinum's complex anatomic structures and intertype confusion among different mediastinal neoplasm pathologies severely hinder accurate diagnosis. To solve these difficulties, we organised a multicentre national collaboration on the basis of privacy-secured federated learning and developed CAIMEN, an efficient chest CT-based artificial intelligence (AI) mediastinal neoplasm diagnosis system.

Methods In this multicentre cohort study, 7825 mediastinal neoplasm cases and 796 normal controls were collected from 24 centres in China to develop CAIMEN. We further enhanced CAIMEN with several novel algorithms in a multiview, knowledge-transferred, multilevel decision-making pattern. CAIMEN was tested by internal (929 cases at 15 centres), external (1216 cases at five centres and a real-world cohort of 11 162 cases), and human–AI (60 positive cases from four centres and radiologists from 15 institutions) test sets to evaluate its detection, segmentation, and classification performance.

Findings In the external test experiments, the area under the receiver operating characteristic curve for detecting mediastinal neoplasms of CAIMEN was 0·973 (95% CI 0·969–0·977). In the real-world cohort, CAIMEN detected 13 false-negative cases confirmed by radiologists. The dice score for segmenting mediastinal neoplasms of CAIMEN was 0·765 (0·738–0·792). The mediastinal neoplasm classification top-1 and top-3 accuracy of CAIMEN were 0·523 (0·497–0·554) and 0·799 (0·778–0·822), respectively. In the human–AI test experiments, CAIMEN outperformed clinicians with top-1 and top-3 accuracy of 0·500 (0·383–0·633) and 0·800 (0·700–0·900), respectively. Meanwhile, with assistance from the computer aided diagnosis software based on CAIMEN, the 46 clinicians improved their average top-1 accuracy by 19·1% (0·345–0·411) and top-3 accuracy by 13·0% (0·545–0·616).

Interpretation For mediastinal neoplasms, CAIMEN can produce high diagnostic accuracy and assist the diagnosis of human experts, showing its potential for clinical practice.

Funding National Key R&D Program of China, National Natural Science Foundation of China, and Beijing Natural Science Foundation.

Introduction

Mediastinal neoplasms are typical thoracic diseases, with around 0·77–1·68% prevalence in the population,[1–3] indicating around 60–130 million patients worldwide. The pathological compartments of mediastinal neoplasms are all-encompassing, ranging from benign masses to malignant tumours.[4] In the past 10 years, the incidence of mediastinal neoplasms has increased,[5] and patients with malignancies might suffer from poor prognosis.[6] Therefore, accurate and timely diagnosis is essential for better treatments and personalised health care in clinical practice.

In the past two decades, CT screening has been gaining popularity as a non-invasive diagnostic solution for various diseases. However, CT-based diagnosis of mediastinal neoplasms remains challenging in practice. First, mediastinal neoplasms and normal structures in the mediastinum can be easily confused due to their appearance similarity and spatial adjacency, making it difficult to exhaustively detect and precisely locate

Lancet Digit Health 2023; 5: e560–70

*Contributed equally

Department of Thoracic Oncology and Surgery, China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China (H Liang PhD, Z Yan BS, Jun Liu PhD, X Xu PhD, W Shao PhD, S Li PhD, Prof W Liang PhD, W Wang PhD, F Cui PhD, H He PhD, Chao Yang PhD, L Jiang PhD, H Wang MD, Prof J He PhD, M Zhang MD); **School of Software, Beijing National Research Center for Information Science and Technology, Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing, China** (R Tang PhD, F Xu PhD); **Institute for Brain and Cognitive Sciences, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China** (Y Guo PhD); **Institute for Brain and Cognitive Sciences, Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China** (Prof Q Dai PhD); **Department of Thoracic Surgery** (Zhigang Li PhD, Z Liu MD) **and Department of Radiology** (H Yu MD), **Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; Department of Thoracic Surgery, The First Affiliated Hospital, School of Medicine, Zhejiang University,**

Hangzhou, Zhejiang, China (X Lin PhD, Prof J Hu PhD, Jiangang Zhao MD)**; Guangdong Association of Thoracic Disease, Guangzhou, China** (Z Yan, C Guo MD, H Zhang MD)**; Department of Radiology, The Second Affiliated Hospital of Guangzhou Medical University, Guangzhou, China** (Huai Chen MD)**; School of Information Science and Technology, Fudan University, Shanghai, China** (Z Gao PhD)**; Department of Physiology, School of Basic Medical Sciences, Guangzhou Medical University, Guangzhou, China** (L Zhao PhD)**; Department of Thoracic Surgery, Lanzhou University Second Hospital, Lanzhou University Second Clinical Medical College, Lanzhou, China** (B Li PhD, C Yin MD, W Mao MD)**; Department of Thoracic Surgery, Gaozhou People's Hospital, Gaozhou, China** (W Lin MD, Y Xie MD)**; Department of Thoracic Surgery, Peking University Shenzhen Hospital, Shenzhen, China** (Jixian Liu MD, X Li PhD, D Wu MD, Q Hou MD)**; Department of Thoracic Surgery, The Second Affiliated Hospital of Soochow University, Suzhou, China** (Y Chen PhD, Y Xue MD)**; Department of Thoracic Surgery, Zhongshan Hospital Fudan University, Shanghai, China** (D Chen PhD)**; Department of Cardiothoracic Surgery, Zhongshan City People's Hospital, Zhongshan, China** (Y Liang MD, W Tang MD)**; Department of Respiratory Medicine, The Second Hospital of Dalian Medical University, Dalian, China** (Prof Qi Wang PhD, E Li MD)**; Department of Thoracic Surgery, Cancer Hospital of Dalian University of Technology, Liaoning Cancer Hospital & Institute, Shenyang, China** (H Liu MD, G Wang MD, P Yu MD)**; Department of Thoracic Surgery, Fujian Medical University Union Hospital, Fuzhou, China** (Prof C Chen PhD, B Zheng PhD, Hao Chen MD)**; Department of Thoracic Surgery, Qingdao Municipal Hospital, University of Health and Rehabilitation Sciences, Qingdao, China** (Zhe Zhang MD, Lunqing Wang MD, A Wang MD, Zongqi Li MD)**; Department of**

Research in context

Evidence before this study

We searched PubMed and Google Scholar for artificial intelligence (AI)-based studies on CT for the diagnosis of mediastinal neoplasms published before Oct 12, 2022, using the search terms "deep learning", "federated learning", "artificial intelligence", or "mediastinal neoplasm", "mediastinal tumor", "mediastinal mass", "thymoma". No study reported AI-based mediastinal neoplasm detection and segmentation outcomes. In the mediastinal neoplasm classification task, only a few diagnosis studies with a small sample size, few participating centres, and few types of diseases have been reported. A study found that the energy spectrum CT parameters might have clinical value in the differential diagnosis of thymic epithelial tumours and thymic cysts, with the area under the receiver operating characteristic curve (AUROC) of 0·978 in the arterial phase. Another single-centre study revealed that the specific PET–CT-based radiomic features with image variables could predict thymoma risk groups. A pilot study also showed that a machine learning-based model would provide a potential tool to facilitate the differential diagnosis of anterior mediastinal cysts and type B1 and B2 thymomas on the basis of contrast chest CT. The AUROC, sensitivity, and specificity of the model were 0·899, 84·6%, and 87·5%. No large-scale and multicentre study focused on multitype mediastinal tumour detection and classification.

Added value of this study

This study reported a chest CT-based AI system for accurate mediastinal neoplasm diagnosis. To our knowledge, we are the first to: (1) explore the whole diagnosis process of mediastinal neoplasm; (2) provide a large-scale multicentre cohort by a 24-centre national collaboration; (3) develop an accurate and robust AI-based system for mediastinal neoplasms; (4) protect privacy by federated learning with confidential and encrypted data sharing; and (5) improve the performance of clinicians in practical clinical use. The successful practice for developing expert-level AI systems in a privacy-secured manner provides an example for promoting the research and application of AI-aided health care.This study also made several technical contributions that other AI-based diagnosis systems can draw lessons from. First, a multiview fusion algorithm mimicking radiologists' reading strategy to comprehensively analyse CT for better figuring out lesions. Second, a transfer learning method to enhance diagnosis capacity on plain CT by incorporating the knowledge in contrast CT. Third, a binary classification decision tree and a multigrain deep learning algorithm to establish the relationship among different pathological types for clinically understandable decision path and accurate type classification.

Implications of all the available evidence

The challenges of mediastinal neoplasm diagnosis have attracted more attention from physicians and researchers. Benefited from its accuracy and robustness, CAIMEN can serve as a useful aid for radiologists from different countries and regions worldwide to precisely diagnose mediastinal neoplasm. We will publish our system after the publication of this work, enabling radiologists anywhere in the world to use it for free. The principles of this system can be applied to building computer aided diagnosis systems for a wide range of diseases. We believe this work can attract broad-spectrum attention and be potentially highly impactful.

mediastinal neoplasms from plain CT images.[7] Second, within mediastinal neoplasms, even experienced radiologists can find it difficult to diagnose the exact pathological types because of the many different pathological types, some of which can be easily confused.[8,9] To overcome those challenges, novel technologies for accurate mediastinal neoplasm detection and classification are highly necessary for clinical practice and research.

In the past ten years, advances in deep learning technologies developed for health care, especially for medical image analysis,[10,11] indicate promising possibilities for the diagnosis of mediastinal neoplasms from plain CT images. However, several technical difficulties are hindering the practical development of deep learning for this task. First, deep learning requires a large amount of data for training, but there is no large-scale cohort for mediastinal neoplasms due to privacy legislation that restricts medical data sharing between multiple centres.[12,13] Second, due to the unclear visual boundaries or clues of mediastinal neoplasms and complicated context information, it is difficult to build an accurate lesion segmentation model for mediastinal neoplasms in plain CT images. Third, due to the homogeneity among fine-grained types of mediastinal neoplasms, it is hard to develop an accurate classification model.

In this work, we aimed to address the above difficulties by constructing CAIMEN, a chest CT-based artificial intelligence (AI) mediastinal neoplasm diagnosis system. We initiated the National Mediastinal Neoplasms Collaboration (NMNC) with 24 well known medical centres from 15 provinces in China and constructed, as far as we know, the first distributed, large-scale, multicentre, and well annotated imaging cohort, including mediastinal neoplasms and normal controls. On the basis of the cohort, we developed CAIMEN by nationwide federated learning with confidential and encrypted data sharing to achieve high diagnostic accuracy. Besides, a multiview fusion algorithm, mimicking the reading strategy of expert radiologists, was proposed for comprehensive CT analysis for better diagnosis. Furthermore, a transfer learning method, enhancing CAIMEN for plain CT images by using learned lesion features from contrast CT images, was proposed to improve lesion detection and segmentation accuracy. Moreover, to address the

difficulties in disease type classification, we proposed a multilevel decision-making strategy on the basis of a binary classification decision tree (BCDTree) and a multigrain deep learning algorithm, leading to a clinically understandable decision path and accurate type classification. CAIMEN covers the whole diagnosis process of mediastinal neoplasm, including the detection, segmentation, and classification covering relatively complete pathological types, with accurate and robust performance in practice. We further developed computer aided diagnosis (CAD) software for clinical use (video), with which the diagnosis performance of clinicians was significantly improved.

Methods

Data preparation

In the NMNC, we collected data from the 24 centres following the same protocol (table, appendix p 3). The 24 centres were divided into 15 internal centres for developing and testing CAIMEN, five external centres for evaluating CAIMEN's generalising performance, and four human–AI centres for comparing the capacity of CAIMEN and human experts (appendix p 4).

In internal centres (appendix p 5), we collected 5160 contrast CT volumes and 5790 plain CT volumes with mediastinal neoplasms and 525 plain normal controls without mediastinal neoplasms from Jan 1, 2010, to Oct 31, 2020. In the detection and segmentation tasks, radiologists annotated 687 contrast CT volumes, which were divided into contrast training (468 volumes) and validation (219 volumes) sets, and annotated 1106 plain CT volumes, which were divided into internal training (581 volumes), validation (159 volumes, 159 normal controls), and test (366 volumes, 366 normal controls) sets. In the classification task, 4981 plain CT volumes were selected and divided into internal training (2974 volumes), validation (1078 volumes), and test (929 volumes) sets. There is no patient overlapping between any training and test sets for different tasks.

In external centres (appendix p 5), we collected 1527 plain CT volumes with mediastinal neoplasms and 271 plain normal controls from Jan 1, 2010, to Oct 31, 2020. In the detection and segmentation tasks, 271 plain CT volumes annotated by radiologists and 271 normal controls were used as external test sets. In the classification task, 1228 plain CT volumes were selected as external test sets.

In human–AI centres (appendix p 5), we collected 142 plain CT volumes with mediastinal neoplasms from Jan 1 to Dec 31, 2021. We selected 60 plain CT volumes as the human–AI test sets for the human–AI test experiment. The data characteristics in different tasks were analysed and summarised (appendix pp 6–7). All data were precisely annotated (appendix pp 8–9). All data were deidentified to protect the patients' privacy.

To further evaluate the generalisation performance of CAIMEN, we used data from the National Lung Screening Trial (NLST)[14] obtained from the National Cancer Data Access System of the US National Cancer Institute (NCI), through a data transfer agreement between the authors and the NCI (project number 868). We collected 26253 cases with plain CT images via the official transfer tool. Because there is no official diagnosis related to mediastinal neoplasms, we selected 11162 participants with the official diagnosis of no significant abnormalities for all three screening times as the NLST test set.

	Participants (n=7825)
Age, years	
≤19	298 (3·81%)
20–39	1751 (22·38%)
40–59	3630 (46·39%)
60–79	2095 (26·77%)
≥80	43 (0·55%)
Unknown	8 (0·10%)
Sex	
Male	3883 (49·62%)
Female	3942 (50·38%)
Risk	
Benign	5237 (66·93%)
Malignant	2304 (29·44%)
Unknown	284 (3·63%)
Pathology	
Thymoma	2388 (30·52%)
Benign cyst	1828 (23·36%)
Thymic carcinoma	527 (6·73%)
Germ cell tumour	455 (5·81%)
Neuroendocrine tumour	113 (1·44%)
Thymic hyperplasia	332 (4·24%)
Lymphoma	279 (3·57%)
Lymphadenosis	213 (2·72%)
Ectopic thyroid gland	98 (1·25%)
Granulomatous inflammation	72 (0·92%)
Neurogenic tumour	785 (10·03%)
Other soft tissue tumour	301 (3·85%)
Metastasis	117 (1·50%)
Unknown	317 (4·05%)

Data are n (%).

Table: **Patient characteristics in the cohort**

The multicentre study was approved by the Ethics Committee of the National Center for Respiratory Medicine/The First Affiliated Hospital of Guangzhou Medical University (Oct 12, 2020; Institutional Review Board number: 2020 No.138). All image data in the NMNC database were deidentified to protect the patients' privacy.

Architecture of CAIMEN

CAIMEN covered the whole diagnosis process including the detection, segmentation, and classification of

Thoracic Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China (J Fu MD, G Zhang MD, J Zhang MD, B Liu MD); **Department of Chest Surgery, Affiliated Cancer Hospital & Institute of Guangzhou Medical University, Guangzhou, China** (Jian Zhao MD); **Department of Thoracic Surgery, Central People's Hospital of Zhanjiang, Zhanjiang, China** (B Deng MD); **Division of Thoracic Surgery, Sichuan Cancer Hospital & Institute, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China** (Y Han PhD, X Leng PhD, Zhiyu Li MD); **Department of Thoracic Surgery, The Affiliated Hospital of Inner Mongolia Medical University, Hohhot, China** (M Zhang, C Liu MD); **Department of Thoracic Surgery, The Third Affiliated Hospital of Chongqing Medical University, Chongqing, China** (T Wang MD, Z Luo MD); **National Cancer Center, National Clinical Research Center for Cancer, Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China** (Chenglin Yang MD, X Guo MD, K Ma MD, Lixu Wang MD); **Department of Thoracic Surgery, The Fourth Affiliated Hospital of China Medical University, Shenyang, China** (W Jiang MD, X Han MD, Qing Wang MD); **Department of Thoracic Surgery, The Third People's Hospital of Shenzhen, Shenzhen, China** (K Qiao MD, Z Xia MD, S Zheng MD); **Department of Thoracic Surgery** (C Xu MD, S Wu MD) **and Department of Radiology** (J Peng MD), **Ganzhou People's Hospital, Ganzhou, China; Department of Cardiothoracic Surgery, Jieyang People's Hospital, Jieyang, China** (Zhifeng Zhang MD, H Huang MD); **Department of Thoracic Surgery, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China** (D Pang MD, Q Liu BSN, J Li MD, X Ding MD); **Department of Thoracic Surgery, The Second Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang, China** (X Liu MD); **Department of**

Radiology, Huizhou First People's Hospital, Huizhou, China (L Zhong MD); **School of Computer Science and Engineering, Sun Yat-sen University, National Supercomputer Center, Guangzhou, China** (Y Lu MD); **Hangzhou Zhuoxi Institute of Brain and Intelligence, Hangzhou, China** (Y He PhD, X Chen ME)

Correspondence to:
Dr Feng Xu, School of Software, Beijing National Research Center for Information Science and Technology, Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing 100084, China
feng-xu@tsinghua.edu.cn

or

Prof Qionghai Dai, Institute for Brain and Cognitive Sciences, Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
daiqh@tsinghua.edu.cn

or

Prof Jianxing He, Department of Thoracic Oncology and Surgery, China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, China
drjianxing.he@gmail.com

See Online for video

See Online for appendix

mediastinal neoplasms (figure 1C). The whole system was developed via a federated learning (appendix p 10) framework to establish connections between the centres within the NMNC with confidential and encrypted data sharing (figure 1B).

In the detection and segmentation tasks, we proposed the multiview fusion algorithm (appendix p 11) mimicking the reading strategy of radiologists to comprehensively analyse the mediastinum areas from the axial, coronal, and sagittal directions of the CT volumes to acquire richer context information for better lesion detection and segmentation. In addition, we used transfer learning (appendix p 12) to transfer learned lesion knowledge from contrast CT images in which mediastinal neoplasms were more distinguishable to CAIMEN for better identifying mediastinal neoplasms in plain CT images. We used TransUNet[15] for predictions in a two-dimensional pattern and finally got a segmentation mask and a detection confidence value (appendix p 13).

In the classification task, we proposed a classification model called Feature Pyramid Transformer Network (FPTN), which accepted a CT volume with its segmentation mask and output the classification confidences (appendix p 14). Besides, we used our large-scale cohort to explore a relationship between different pathological types by recursively grouping together those with larger similarities in a data-driven pattern and finally constructed a BCDTree (figure 1C; appendix p 15). Then multiple FPTN focusing on only a subgroup of pathological types jointly determined the classification prediction of each CT volume. In this way, each classification module's learning difficulty was significantly reduced, finally improving the classification accuracy of CAIMEN.

Statistical analysis

For evaluations, we used the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, precision, and F1 score (a harmonic mean of precision and recall) in the detection task, dice score in the segmentation task, and top-1 and top-3 accuracy and AUROC in the classification task (appendix p 16). For comparisons, we used the one-tailed Delong test approach[16] for AUROC and the one-tailed Wilcoxon signed rank test with continuity correction[17] for other metrics. In particular, when comparing federated

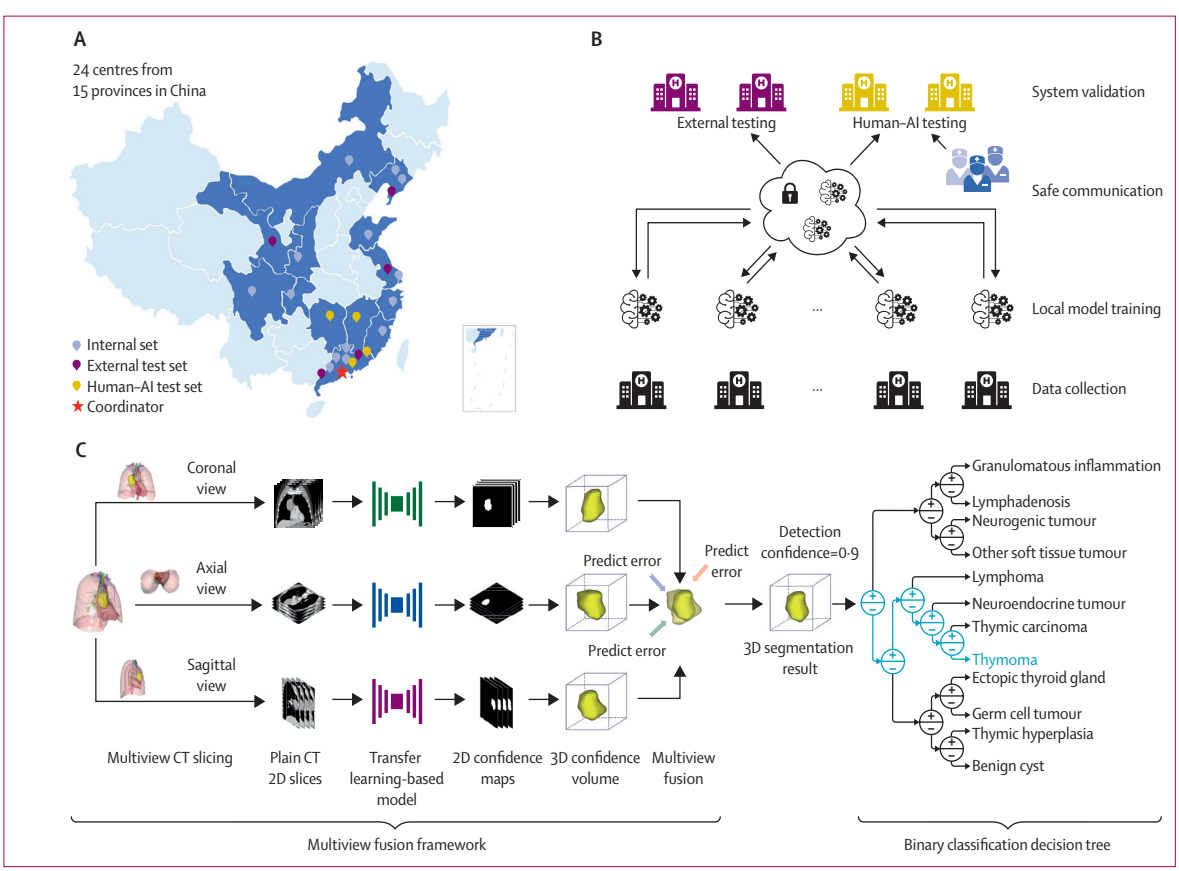

Figure 1: **Overview of CAIMEN**
(A) The centre distribution of the National Mediastinal Neoplasms Collaboration. (B) Federated learning framework and test process of CAIMEN. (C) The whole system architecture of CAIMEN. AI=artificial intelligence. CAIMEN=a chest CT-based AI mediastinal neoplasm diagnosis system. 2D=two-dimensional. 3D=three-dimensional.

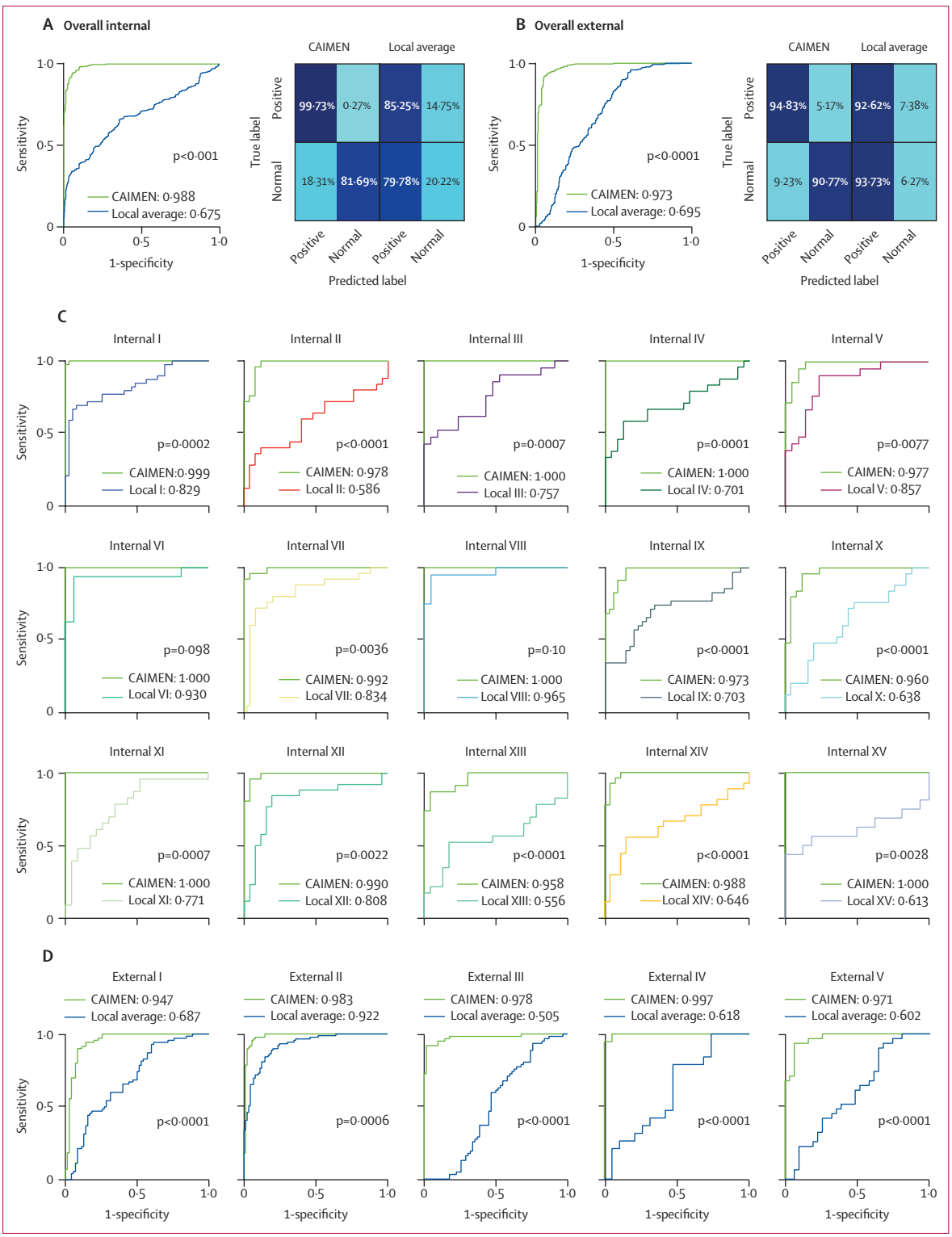

Figure 2: **Detection performance of CAIMEN and local models**
(A) The overall ROC curves and confusion matrices on internal test sets. (B) The overall ROC curves and confusion matrices on external test sets. (C) The ROC curves on each internal test set. (D) The ROC curves on each external test set. CAIMEN=a chest CT-based artificial intelligence mediastinal neoplasm diagnosis system. ROC=receiver operating characteristic curve. Internal I–XV=the 15 internal centres. External I–V=the five external centres.

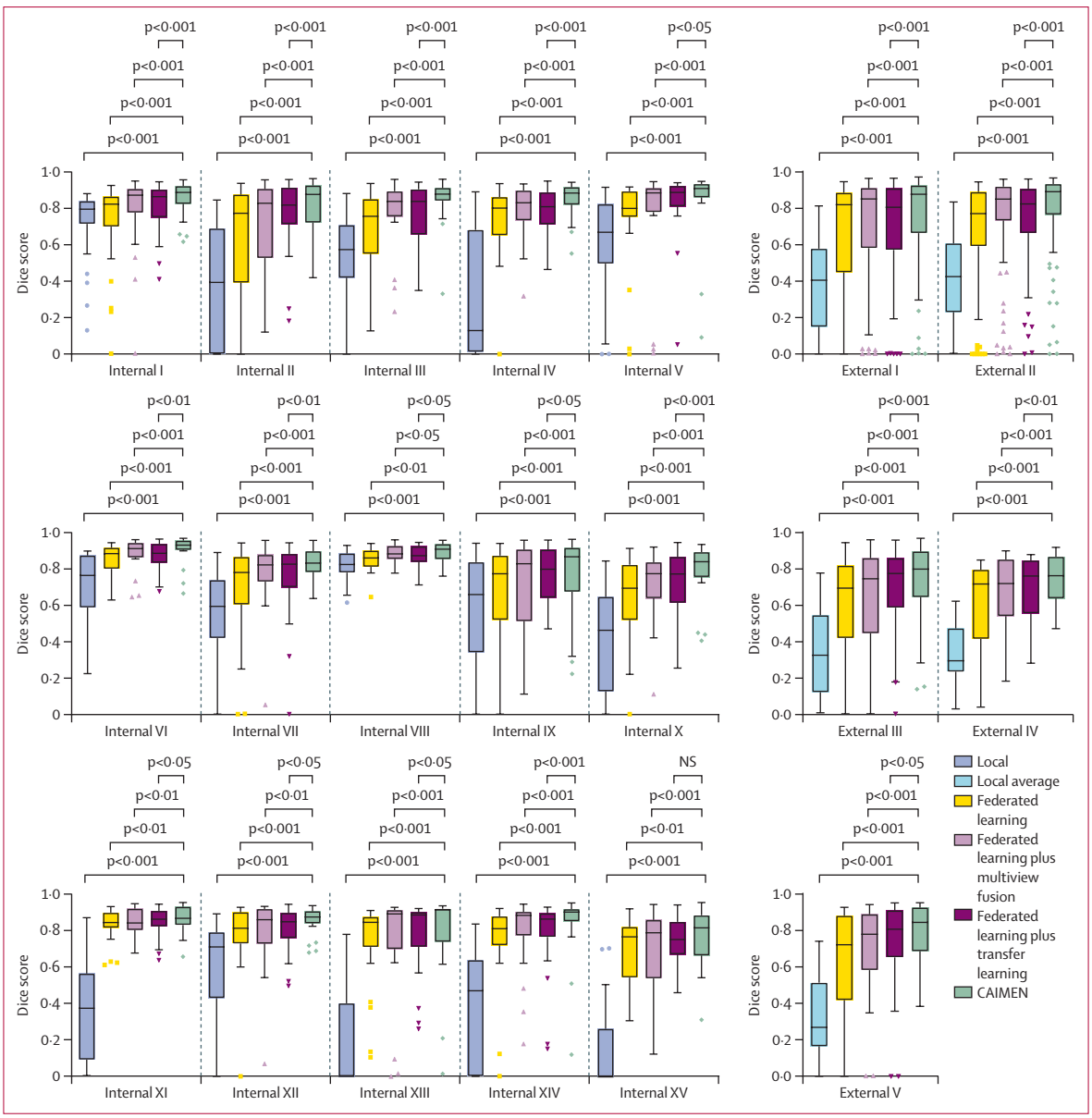

Figure 3: **Segmentation performance of CAIMEN, local models, and federated learning models integrated with different techniques**
The dice scores on each internal test set and external test set. Error bars in the box-plot elements represent Tukey Confidence Interval. CAIMEN=a chest CT-based artificial intelligence mediastinal neoplasm diagnosis system. NS=not significant. Internal I–XV=the 15 internal centres. External I–V=the five external centres.

learning and centralised learning, we used two-tailed versions of the methods mentioned above.

We deemed p values less than 0·05 to be statistically significant. The 95% CIs of all metrics were computed through bootstrapping. Graphpad Prism 6.0 software (La Jolla, California, USA) and BioRender were used for drawing figures.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, writing of the manuscript, or the decision to submit.

Results

CAIMEN had high accuracy for mediastinal neoplasm detection (figure 2). The evaluation for detection on internal test sets yielded an average AUROC of 0·988 (95% CI 0·983–0·993), sensitivity of 0·997 (0·989–1·000), and specificity of 0·817 (0·779–0·855). On external test sets, CAIMEN had an average AUROC of 0·973 (0·969–0·977), sensitivity of 0·948 (0·941–0·955), and specificity of 0·908 (0·898–0·916).

We then used the NLST test set to evaluate the cross-race detection capacity. Although all samples here had an official diagnosis of no significant abnormalities,

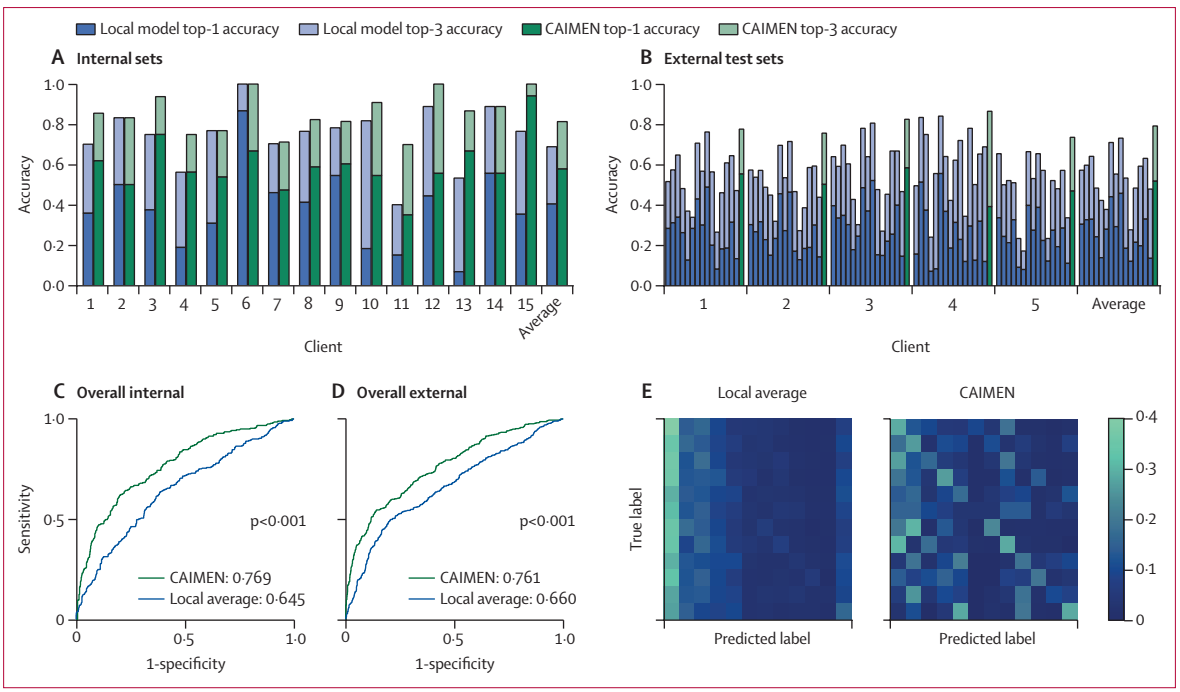

Figure 4: **Classification performance of CAIMEN and local models**
(A) The top-1 and top-3 accuracies on internal test sets. (B) The top-1 and top-3 accuracies on external test sets. (C) The overall benign–malignant ROC curves on internal test sets. (D) The overall benign–malignant ROC curves on external test sets. (E) The overall top-3 accuracy confusion matrices on external test sets. CAIMEN=a chest CT-based artificial intelligence mediastinal neoplasm diagnosis system. ROC=receiver operating characteristic curve.

CAIMEN detected 13 cases confirmed by radiologists to contain mediastinal neoplasms (appendix p 17), which showed CAIMEN's excellent detection capacity.

Integrated with federated learning, multiview fusion, and transfer learning, CAIMEN finally had a much higher detection performance than local models without any of the techniques (figure 2). On internal test sets, compared with local models, CAIMEN improved the average AUROC by 46·4% (0·675 [95% CI 0·632–0·714] for local models *vs* 0·988 [0·983–0·993] for CAIMEN; p<0·001), sensitivity by 17·0% (0·852 [0·814–0·888] for local models *vs* 0·997 [0·989–1·000] for CAIMEN; p<0·001), and specificity by 304·5% (0·202 [0·161–0·246] for local models *vs* 0·817 [0·779–0·855] for CAIMEN; p<0·001). On external test sets, CAIMEN increased local models' average AUROC by 40·0% (0·695 [0·649–0·738] for local models *vs* 0·973 [0·958–0·985] for CAIMEN; p<0·001), sensitivity by 2·4% (0·926 [0·893–0·956] for local models *vs* 0·948 [0·923–0·970] for CAIMEN; p<0·001), and specificity by 1341·3% (0·063 [0·037–0·092] for local models *vs* 0·908 [0·875–0·941] for CAIMEN; p<0·001). Subgroup analysis also showed CAIMEN's consistent advantages over local models (appendix p 18).

In the segmentation task, CAIMEN could accurately identify the lesion areas in plain CT images (figure 3). On internal test sets, CAIMEN had an average dice score of 0·838 (95% CI 0·823–0·854). On external test sets, CAIMEN had an average dice score of 0·765 (0·738–0·792). For different mediastinal neoplasm types, CAIMEN performed well and predicted accurate segmentation results (appendix p 19).

CAIMEN significantly improved the average dice score of local models (figure 3), by 67·9% on internal test sets (0·499 [95% CI 0·465–0·536] for local models *vs* 0·838 [0·823–0·854] for CAIMEN; p<0·001) and 105·1% on external test sets (0·373 [0·346–0·400] for local models *vs* 0·765 [0·738–0·792] for CAIMEN; p<0·001). In the subgroup analysis, CAIMEN also consistently outperformed local models (appendix p 20).

The multiview fusion and transfer learning algorithms could further improve the detection and segmentation accuracy on the basis of the federated learning technique. In the detection task (appendix p 21), multiview fusion increased the average AUROC by 5·2% on internal test sets (0·925 [95% CI 0·904–0·944] for federated learning *vs* 0·973 [0·962–0·982] for federated learning plus multiview fusion; p<0·001) and 7·4% on external test sets (0·877 [0·846–0·904] for federated learning *vs* 0·942 [0·922–0·961] for federated learning plus multiview fusion; p<0·001). Transfer learning improved the average AUROC by 3·8% on internal test sets (0·925 [0·904–0·944] for federated learning *vs* 0·960 [0·946–0·971] for federated learning plus transfer learning; p<0·001) and 6·6% on external test sets (0·877 [0·846–0·904] for federated learning *vs* 0·935 [0·914–0·955] for federated learning plus transfer learning; p<0·001).

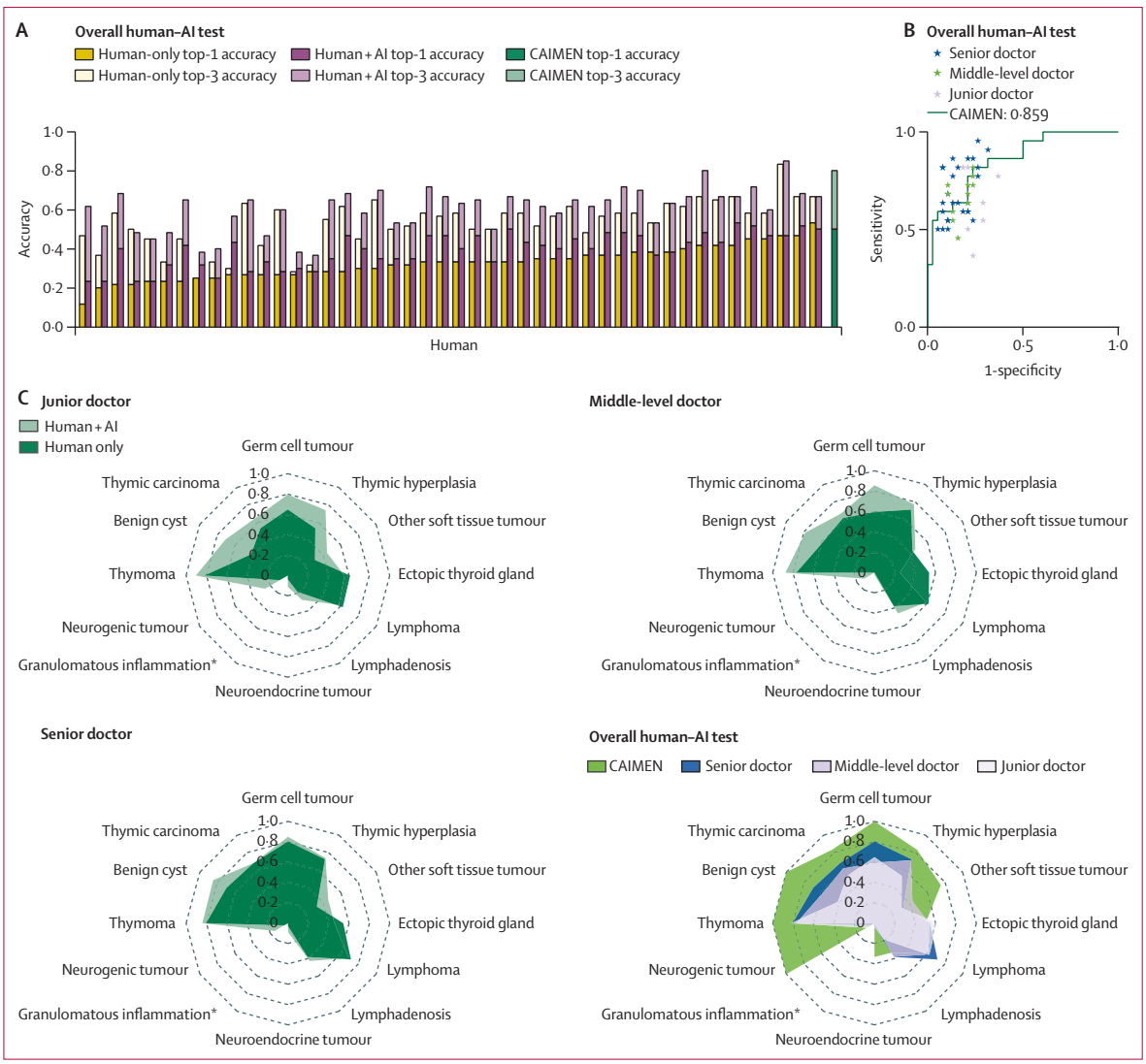

Figure 5: **Human–AI test experiments**
(A) The overall classification accuracy of CAIMEN and doctors with or without CAIMEN's assistance. (B) The overall benign-malignant ROC curves of CAIMEN and doctors with different qualifications. (C) The top-3 accuracy of doctors with different qualifications with or without CAIMEN's assistance for each pathological type. AI=artificial intelligence. CAIMEN=a chest CT-based AI mediastinal neoplasm diagnosis system. ROC=receiver operating characteristic curve. *We collected no samples of granulomatous inflammation in the human–AI validation sets.

In the segmentation task (figure 3), multiview fusion improved average dice score by 6·4% on internal test sets (0·733 [0·711–0·757] for federated learning *vs* 0·780 [0·759–0·802] for federated learning plus multiview fusion; p<0·001) and 10·2% on external test sets (0·636 [0·602–0·671] for federated learning *vs* 0·701 [0·668–0·732] for federated learning plus multiview fusion, p<0·001). Transfer learning increased average dice score by 8·2% on internal test sets (0·733 [0·711–0·757] for federated learning *vs* 0·793 [0·776–0·809] for federated learning plus transfer learning; p<0·001) and 12·6% on external test sets (0·636 [0·602–0·671] for federated learning *vs* 0·716 [0·687–0·744] for federated learning plus transfer learning; p<0·001).

CAIMEN had a good ability to distinguish the 12 pathological types of mediastinal neoplasms (figure 4A, B, E), with average top-1 accuracy of 0·578 (95% CI 0·545–0·609) on internal test sets and 0·523 (0·497–0·554) on external test sets. When allowed to recommend three diagnoses, CAIMEN reached average top-3 accuracy of 0·814 (0·788–0·840) on internal test sets and 0·799 (0·778–0·822) on external test sets. As for classification between benign masses and malignant tumours (figure 4C, D), CAIMEN had an average AUROC of 0·769 (0·738–0·802) on internal test sets and 0·761 (0·733–0·792) on external test sets. The effect of BCDTree is discussed in the appendix (p 22).

Compared with local models, CAIMEN improved average top-1 and top-3 classification accuracy (figure 4A, B) by 43·1% (0·404 [95% CI 0·372–0·439] for local models *vs* 0·578 [0·545–0·609] for CAIMEN; p<0·001) and 18·1% (0·689 [0·661–0·719] for local models *vs* 0·814 [0·788–0·840] for CAIMEN, p<0·001) on internal test sets, respectively, and improved average top-1 and top-3 accuracy by 93·0% (0·271 [0·246–0·295] for local models *vs* 0·523 [0·497–0·554] for CAIMEN; p<0·001) and 51·3% (0·528 [0·501–0·556] for local models *vs* 0·799 [0·778–0·822] for CAIMEN; p<0·001) on external test sets, respectively. When distinguishing benign masses and malignant tumours (figure 4C, D), CAIMEN also outperformed local models in average AUROC by 19·2% (0·645 [0·606–0·683] for local models *vs* 0·769 [0·738–0·802] for CAIMEN; p<0·001) on internal test sets and 15·3% (0·660 [0·627–0·693] for local models *vs* 0·761 [0·733–0·792] for CAIMEN, p<0·001) on external test sets. In the subgroup analysis, CAIMEN showed consistent advantages over local models (appendix p 23).

We developed CAD software for clinical usage (video). Compared with human experts (appendix p 24), CAIMEN outperformed them in average top-1 accuracy by 44·9% (0·345 for human experts *vs* 0·500 for CAIMEN) and top-3 accuracy by 46·8% (0·545 for human experts *vs* 0·800 for CAIMEN). However, with the assistance from the CAD software by visual information and prediction results, human experts' performance was improved in average top-1 accuracy by 19·1% (0·345 without assistance *vs* 0·411 with assistance) and top-3 accuracy by 13·0% (0·545 without assistance *vs* 0·616 with assistance; figure 5C). For benign masses and malignant tumour classification, CAIMEN could have competitive distinguishing capability as compared with human experts with an average of 12·7 years of experience (figure 5B).

In this study, we wanted to identify whether the model performance would be weakened in CAIMEN due to data decentralisation. Therefore, after collecting data from all centres for public release, we developed a centralised learning system with centralised data and compared CAIMEN with centralised learning in different tasks (appendix p 25). As shown by evaluations, there was no significant difference between the two systems' average AUROC on internal test sets (0·987 [95% CI 0·980–0·993] for centralised learning *vs* 0·988 [0·983–0·993] for CAIMEN; p=0·660) and external test sets (0·976 [0·962–0·987] for centralised learning *vs* 0·973 [0·958–0·985] for CAIMEN; p=0·316) in the detection task, dice score on internal test sets (0·833 [0·818–0·851] for centralised learning *vs* 0·838 [0·823–0·854] for CAIMEN; p=0·946) and external test sets (0·760 [0·733–0·787] for centralised learning *vs* 0·765 [0·738–0·792] for CAIMEN; p=0·712) in the segmentation task, and top-3 accuracy on internal test sets (0·809 [0·786–0·835] for centralised learning *vs* 0·814 [0·788–0·840] for CAIMEN; p<0·001) and external test sets (0·803 [0·781–0·826] for centralised learning *vs* 0·799 [0·778–0·822] for CAIMEN, p<0·001) in the classification task, showing that CAIMEN could not only ensure data privacy protection, but also prevent performance degradation due to data decentralisation.

Discussion

The challenges of mediastinal neoplasm diagnosis have attracted increased attention from physicians and researchers.[18] As far as we know, CAIMEN is the first clinically relevant AI diagnosis system for mediastinal neoplasms via a large-scale multicentre project. On average, CAIMEN took around 15 s to diagnose one CT volume with high accuracy of lesion localisation, segmentation, and classification, showing its potential for providing high-throughput, non-invasive, and low-cost diagnosis for mediastinal neoplasms during clinical practice. The national collaboration further validates the value of federated learning in developing a distributed deep learning-based medical diagnosis system beyond clinicians' observational power and not inferior to centralised learning, especially for rare diseases such as mediastinal neoplasms.

Mediastinal neoplasms often present very irregular shapes and indistinct boundaries, causing difficulties for physicians to precisely segment. As far as we have learned, this study is the first to explore AI-based mediastinal neoplasm segmentation, with satisfactory outcomes. Even in some challenging cases, such as visceral mediastinal neoplasms with complex surrounding anatomy, CAIMEN remains high in segmentation performance (average dice score of 0·720 on external test sets). Small mediastinal neoplasms with maximum diameters smaller than 3 cm are also well segmented (average dice score of 0·620 on external test sets), which provides a good foundation for accurately segmenting small mediastinal neoplasms in clinical practice. In the subgroup analysis based on the pathological type, CAIMEN's performance on lymphadenosis (average 0·493 on external test sets) and neuroendocrine tumours (average 0·505 on external test sets) was poorer than that on other types. This could be due to the difficulty of distinguishing the two types of mediastinal neoplasms with adjacent tissues in plain CT images.

To exhaustively detect mediastinal neoplasms, a mediastinal neoplasm detection subsystem derived from the segmentation subsystem was developed. As shown by current results, CAIMEN has sufficiently high detection sensitivity even for small mediastinal neoplasms with maximum diameters smaller than 3 cm (average 0·829 on external test sets). It is worth noting that although CAIMEN presented relatively poor segmentation performance on lymphadenosis and neuroendocrine tumours, its detection sensitivities remain high enough to meet clinical needs (average 0·816 for lymphadenosis and 0·969 for neuroendocrine

tumours). In addition, the detected and confirmed 13 false-negative cases in the NLST cohort show CAIMEN's potential to help reduce the missed diagnosis rate.

In the mediastinal neoplasm classification task, no AI-assisted diagnostic study has covered diseases other than thymomas and benign cysts, nor has a large-scale study ever developed a clinically applicable diagnostic system for pan-mediastinal neoplasms.[19–21] Our study is the first to cover relatively complete pathological types in the mediastinal neoplasm and develop an AI diagnostic system with high classification accuracy. Besides, the human–AI test experiments also show CAIMEN's superior performance over human experts and its good assisting capacity. We do not include metastases in our study because we considered that symptoms and evidence would be more specific based on image materials from the primary tumours than from the metastases.

The innovative designs and algorithms in CAIMEN might promote other AI-based medical research. First, the multiview fusion algorithm combines the multiview information to reduce prediction errors. This algorithm might be recommended for all three-dimensionally formed data to improve the diagnosis performance of AI. Second, the transfer learning algorithm incorporates features of contrast CT images to improve CAIMEN's performance on plain CT images. This algorithm might also be applied to other medical problems with cross-modality data. Third, the BCDTree establishes a data-driven disease relationship highly consistent with radiologists' knowledge, so we believe it might have great medical significance for finding new medical knowledge.

Several studies have already applied federated learning to construct a robust and generalisable model in clinical practice. However, most of those studies only focused on relatively simple medical problems, such as identification or mortality prediction for diseases.[22,23] In contrast, our study covers the whole diagnosis process of mediastinal neoplasms, including detection, segmentation, and classification. Besides, CAIMEN also considers more realistic cases with extremely different cross-centre data distributions. For example, among the 15 centres involved in the training, only three centres held data on lymphadenosis, and only four centres held data on granulomatous inflammation. By considering and overcoming these realistic and challenging cases, CAIMEN can be more clinically applicable in medical practice.

Future work is needed to further improve the functionality of CAIMEN. In the multiview fusion algorithm, predictions from three directions were fused by averaging their confidence values to reach the best performance among our various explorations. An approach considering more characteristics of the three views might have better performance. In the detection task, we took the maximum confidence value from the fused prediction as the detection confidence achieving the best results. We believe other methods can be developed to increase CAIMEN's detection accuracy. In comparison with centralised learning in the classification task, CAIMEN had relatively lower average top-1 accuracy than centralised learning, presenting limitations of our federated learning algorithms in the case of extremely different data distributions between centres. A federated learning algorithm overcoming this problem might make CAIMEN more applicable in medical practice. Finally, although the BCDTree might be beneficial for knowledge exploration, expensive computing resources are required to construct and train it. A strategy less resource-consuming might further increase its practical application value.

In conclusion, our study explores the first step towards mediastinal neoplasm diagnosis via deep learning and federated learning. We believe this study could promote related research on mediastinal neoplasms. Besides, the developed system might also support physicians for more accurate diagnosis in clinical practice. In addition, we believe the general algorithms in our system could also be applied to other medical problems for more precise disease diagnosis.

Contributors

RT, HeL, YG, FX, QD, and JHe provided the conception and design of this study. JHe, QD, and FX coordinated and organised the research team to complete this work. RT developed the CAIMEN system. YG and FX assisted in improving the system. RT, HeL, YG, and FX provided the design of the experiments. RT did the experimental analysis. HeL did the clinical comparison on the the CAIMEN system with human experts and RT did the outcome analysis. ZhicL, ZhigL, and XLin led the data collection and all other authors participated in the data collection. RT, HeL, YG, and FX wrote the paper. All authors read and approved the paper. All authors had access to all the raw datasets. RT and HeL have accessed and verified the data. RT, HeL, YG, ZhicL, ZhigL, XLin, FX, QD, and JHe were responsible for the decision to submit the manuscript.

Declaration of interests

We declare no competing interests.

Data sharing

The deidentified raw DICOM files for the development and test of CAIMEN and corresponding patients' clinical information can be accessed by sending a request to thss15_tangrj@163.com with detailed reasons for usage. The source code, model parameters, and the CAD software will be available at https://github.com/TangRuijie/caimen after publication. A demo video of the CAD software is available at https://gitlab.com/caimen/demo/.

Acknowledgments

We thank Lindsey Hamblin for editing the manuscript for better readability. We also thank the following individuals for their support in AI and human comparison experiments: Xinjian Chen, Ying Chen, Heng Chu, Shan Gao, Haiyong Gu, Zhongyi Ji, Cong Lan, Chunguang Li, Fei Li, Lun Li, Yongshun Li, Hanghui Liu, Wei Liu, Xiaobin Liu, Qingquan Luo, Fengyuan Peng, Yukai Peng, Dongxue Qin, Wenpin Qiu, Ge Sun, Lin Tan, Dongfei Wang, Bomeng Wu, Xihao Xie, Yu Xiong, and Baoping Zhang. This study is supported by the National Key R&D Program of China (2020AAA0105500 [YG and QD] and 2018YFA0704000 [FX]); National Natural Science Foundation of China (61822111 [FX], 61727808 [FX], 62021002 [FX], 61971260 [YG and QD], and U21B2013 [YG]); Beijing Natural Science Foundation (JQ19015 [FX]); the Zhejiang Provincial Natural Science Foundation (grant LDT23F02024F02 [YHe]) High-level University Construction Project of

Guangzhou Medical University (grants 20182737 [JHe], 201721007 [JHe], 201715907 [JHe], and 2017160107 [JHe]); the Guangdong High Level Hospital Construction "reaching peak" Plan; Guangzhou Institute of Respiratory Health Open Project (funds provided by China Evergrande Group; project numbers 2020GIRHHMS01 [JHe], 2020GIRHHMS09 [JHe], and 2020GIRHHMS10 [JHe]); and Guangzhou Medical University Discipline Construction Funds (basic medicine; JCXKJS2022A11 [JHe]).

Editorial note: The Lancet Group takes a neutral position with respect to territorial claims in published maps and institutional affiliations.

References

1 Henschke CI, Lee IJ, Wu N, et al. CT screening for lung cancer: prevalence and incidence of mediastinal masses. *Radiology* 2006; **239:** 586–90.

2 Yoon SH, Choi SH, Kang CH, Goo JM. Incidental anterior mediastinal nodular lesions on chest CT in asymptomatic subjects. *J Thorac Oncol* 2018; **13:** 359–66.

3 Miyazawa R, Matsusako M, Nozaki T, et al. Incidental mediastinal masses detected at low-dose CT screening: prevalence and radiological characteristics. *Jpn J Radiol* 2020; **38:** 1150–57.

4 Strollo DC, Rosado de Christenson ML, Jett JR. Primary mediastinal tumors. Part 1: tumors of the anterior mediastinum. *Chest* 1997; **112:** 511–22.

5 Roden AC, Fang W, Shen Y, et al. Distribution of mediastinal lesions across multi-institutional, international, radiology databases. *J Thorac Oncol* 2020; **15:** 568–79.

6 Juanpere S, Cañete N, Ortuño P, Martínez S, Sanchez G, Bernado L. A diagnostic approach to the mediastinal masses. *Insights Imaging* 2013; **4:** 29–52.

7 Duwe BV, Sterman DH, Musani AI. Tumors of the mediastinum. *Chest* 2005; **128:** 2893–909.

8 Nakazono T, Yamaguchi K, Egashira R, Mizuguchi M, Irie H. Anterior mediastinal lesions: CT and MRI features and differential diagnosis. *Jpn J Radiol* 2021; **39:** 101–17.

9 Carter BW, Benveniste MF, Marom EM. Diagnostic approach to the anterior/prevascular mediastinum for radiologists. *Mediastinum* 2019; **3:** 18.

10 Zhang K, Liu X, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* 2020; **181:** 1423–33.e11.

11 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; **25:** 30–36.

12 Price WN 2nd, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019; **25:** 37–43.

13 McCall B. What does the GDPR mean for the medical community? *Lancet* 2018; **391:** 1249–50.

14 Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011; **365:** 395–409.

15 Chen J, Lu Y, Yu Q, et al. TransUNet: transformers make strong encoders for medical image segmentation. *arXiv* 2021; published online Feb 8. https://doi.org/10.48550/arXiv.2102.04306 (preprint).

16 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44:** 837–45.

17 Wilcoxon F. Individual comparisons by ranking methods. In: Kotz S, Johnson NL, eds. Breakthroughs in statistics. New York, NY: Springer, 1992: 196–202.

18 Gentili F, Pelini V, Lucii G, et al. Update in diagnostic imaging of the thymus and anterior mediastinal masses. *Gland Surg* 2019; **8** (suppl 3)**:** S188–207.

19 Zhou Q, Huang X, Xie Y, Liu X, Li S, Zhou J. Role of quantitative energy spectrum CT parameters in differentiating thymic epithelial tumours and thymic cysts. *Clin Radiol* 2022; **77:** 136–41.

20 Ozkan E, Orhan K, Soydal C, et al. Combined clinical and specific positron emission tomography/computed tomography-based radiomic features and machine-learning model in prediction of thymoma risk groups. *Nucl Med Commun* 2022; **43:** 529–39.

21 Liu L, Lu F, Pang P, Shao G. Can computed tomography-based radiomics potentially discriminate between anterior mediastinal cysts and type B1 and B2 thymomas? *Biomed Eng Online* 2020; **19:** 89.

22 Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med* 2021; **27:** 1735–43.

23 Bai X, Wang H, Ma L, et al. Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nat Mach Intell* 2021; **3:** 1081–89.