

<https://doi.org/10.1038/s41746-026-02990-y>

Transparent chest radiograph foundation model enables explainable human disease profiling



Chin Lin^{1,2}, Kai-Chieh Chen^{1,2}, Jun-Wei Huang³, Wen-Hui Fang³, Wei-Chou Chang⁴, Kai-Hsiung Ko⁴, Yi-Chih Hsu⁴, Chin-Sheng Lin^{1,2,5}, Shih-Hua Lin⁶ & Dung-Jang Tsai^{1,2}✉

Chest radiography (CXR) is widely accessible, and its ability to capture subtle manifestations of systemic disease remains underexplored. We developed a contrastively pretrained multimodal CXR foundation model using large-scale image–report pairs and evaluated its capacity to predict diverse human diseases. Using 1074 phecodes derived from electronic health records, we trained linear probes on frozen image embeddings and validated performance across three independent cohorts ($n = 90,911$; $n = 79,786$; $n = 60,282$). The model significantly predicted 554 prevalent and 457 incident phenotypes, with 60 prevalent and 42 incident phenotypes demonstrating consistently high discrimination across all datasets. To enhance interpretability, both diseases and 57 radiologist-curated CXR features were embedded into a shared representation space. This co-embedding analysis identified 28 phenotype clusters driven by recognizable imaging patterns, including cardiomegaly, atherosclerosis, and ground-glass opacities, and enabled reconstruction of most predictions with strong explanatory performance ($R^2 \geq 0.85$). Importantly, the embeddings captured imaging signatures associated with near-term cardiovascular events and critical illness. These findings demonstrate that CXR foundation model embeddings encode rich, clinically relevant information beyond conventional interpretation, providing a scalable and interpretable framework for comprehensive disease profiling and a foundation for future prospective validation.

Chest X-ray (CXR) is the world's most common imaging examination¹. It reveals a wide range of thoracic pathologies, making it indispensable for routine diagnosis². Recent advances in deep learning have enabled artificial intelligence (AI) to interpret CXRs with radiologist-level accuracy for many detection tasks³. However, these models depend on large, manually labeled datasets⁴, which are time-consuming and expensive to create⁵, and they can only recognize conditions explicitly annotated during training^{6,7}. As a result, conventional AI struggles to capture subtle clues to rare or future diseases beyond its predefined labels, limiting the scope of CXR analysis.

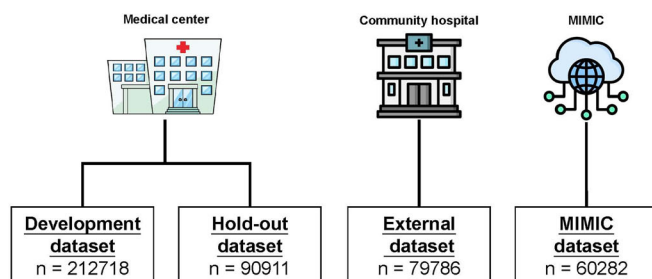
Emerging foundation models in medical imaging address these limitations by using self-supervised learning on massive CXR–report pairs to learn general-purpose image representations^{8–10}. Once pretrained on millions of studies, they can perform zero-shot predictions: ingesting new

natural-language labels and predicting the corresponding findings without additional annotations^{8–10}. Yet their zero-shot capability is confined to features mentioned in the reports, leaving emerging applications, such as predicting mortality^{11,12}, physiologic age^{13,14}, and future major adverse cardiovascular events¹⁵, out of reach. Fortunately, subsequent work has shown that foundation-model embeddings, combined with simple linear probes and minimal labeled data, can accurately predict diseases not specified in the pre-trained dataset¹⁶. To date, however, no study has systematically explored AI–CXR associations across a broad phenotype spectrum, a gap that our work aims to fill.

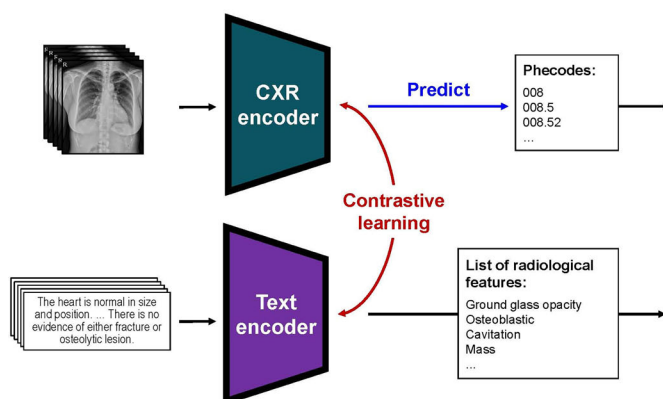
Despite notable gains in disease detection^{17,18}, AI–CXR's clinical adoption is hindered by poor interpretability—a characteristic “black-box” trait of modern AI¹⁹—that makes large-scale implementation challenging^{20,21}. This

¹Medical Technology Education Center, School of Medicine, College of Medicine, National Defense Medical University, Taipei, Taiwan, R.O.C.. ²Military Digital Medical Center, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.. ³Department of Family and Community Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.. ⁴Department of Radiology, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.. ⁵Division of Cardiology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.. ⁶Division of Nephrology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, R.O.C.. ✉e-mail: oo800217@gmail.com

a. Dataset generation



b. Model training



c. Result interpretation

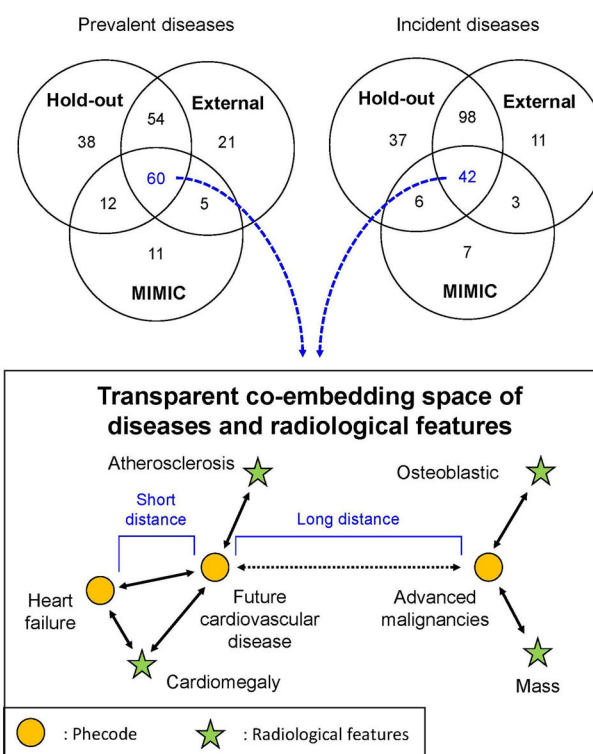


Fig. 1 | Study design, multimodal model training, and disease–feature co-embedding interpretation. **a** Chest radiographs and linked electronic health record data were drawn from three sources: Development dataset (academic medical center; no. of patients = 212,718); Hold-out dataset (academic medical center; no. of patients = 90,911); External dataset (community hospital; no. of patients = 79,786); and MIMIC dataset (public dataset; no. of patients = 60,282). We only selected one chest radiograph at random from each patient. **b** A contrastive learning framework jointly embeds image and text domains. A convolutional encoder ingests chest X-rays (CXR), and a transformer-based text encoder processes radiology report

sentences. During training, paired CXR and report features are aligned in a shared latent space, while the image branch is simultaneously supervised to predict phecde labels via a classification head. **c** Venn diagrams summarize the number of prevalent (left) and incident (right) phecodes meeting performance thresholds. The intersection highlights phecodes consistently high-performing across all three datasets. Below is a section of the learned co-embedding space that visualizes distances between specific phecodes (yellow circles) and radiological features (green stars), where shorter distances indicate stronger image–text associations.

creates a curious paradox: although radiologists assisted by AI-CXR achieve higher diagnostic accuracy than those without AI-CXR support, their performance still falls significantly short of the AI-CXR system alone²²—a pattern that has been observed across other medical applications as well²³. Existing explainability methods, like saliency maps, localize attention but not the underlying features^{24,25}, and fail to capture fine geometric details. Moreover, AI often latches onto confounding “shortcuts”—inferring age, sex, or device presence—and uses these spurious cues for prediction, compromising generalizability²⁶. A truly interpretable AI should identify which classic radiological signs drive its predictions, localize and quantify lesions, and then map those quantified features to multiple diseases.

Figure 1 provides an overview of our study design, multimodal model training, and co-embedding interpretation workflow. We leveraged four independent cohorts, randomly selecting one CXR per patient. Using a contrastive learning-based foundation with a linear probing method to predict 1074 phecodes. Three validation datasets identify 60 prevalent and 42 incident phecodes with consistently high performance. Finally, by analyzing learned embeddings for both diseases and radiological features, we reveal a transparent co-embedding landscape in which shorter distances indicate the specific image findings that drive each disease prediction. Together, this pipeline enables scalable, interpretable disease profiling from routine chest radiographs. Rather than emphasizing the breadth of individual phenotype predictions, this study focuses on whether diverse systemic diseases converge onto a limited set of shared radiographic signatures within a unified embedding space. We hypothesized that large-scale phenotype

profiling would reveal structured clustering driven by classical imaging features rather than isolated disease-specific signals.

Results

Study population

Table 1 presents baseline demographics for the development ($n = 212,718$), hold-out ($n = 90,911$), external ($n = 79,786$), and MIMIC ($n = 60,282$) cohorts. In the development and hold-out datasets, ~59% of CXRs were performed in the emergency or inpatient setting; this was 48.9% in the external dataset and 100% in the MIMIC dataset. Poster-anterior (PA) views comprised 94.2% of the development and hold-out studies, 95.5% of external, but only 58.3% of MIMIC. Patients in the first three datasets were all Asian, while only 4.1% of MIMIC were; gender was balanced (~48–52% female). Mean age ranged from 58.1 to 62.8 years. Development and hold-out patients averaged 7.0 prior diagnoses (external 12.2; MIMIC 7.4 ± 13.3), and 3.2, 3.1, 3.0, and 7.4 comorbidities, respectively.

Phenotype prediction performance

A total of 1074 phecodes met our inclusion criteria and were analyzed for both prevalent and incident outcomes. Supplementary Data 1 and 2 report, for each phecde, the case counts (prevalent) or event counts (incident) by dataset. Supplementary Data 3 and 4 summarize, for all 1074 prevalent and 1,074 incident phecodes, respectively, the discrimination achieved by (a) CXR embeddings alone, (b) demographic information, (c) CXR embeddings plus imaging view, acquisition position, patient race/ethnicity, gender,

Table 1 | Patient characteristics of each dataset

	Development dataset (<i>n</i> = 212,718)	Hold-out dataset (<i>n</i> = 90,911)	External dataset (<i>n</i> = 79,786)	MIMIC dataset (<i>n</i> = 60,282)
<i>Position</i>				
Emergency room or Inpatient department	123,963 (59.0%)	53,034 (59.0%)	38,492 (48.9%)	60,282 (100.0%)
Outpatient department	86,286 (41.0%)	36,808 (41.0%)	40,218 (51.1%)	0 (0.0%)
<i>Imaging type</i>				
PA view	200,373 (94.2%)	85,647 (94.2%)	76,163 (95.5%)	35,120 (58.3%)
AP view	12,345 (5.8%)	5,264 (5.8%)	3,623 (4.5%)	25,162 (41.7%)
<i>Race</i>				
Asian	212,718 (100.0%)	90,911 (100.0%)	79,786 (100.0%)	2,450 (4.1%)
White	0 (0.0%)	0 (0.0%)	0 (0.0%)	37,722 (62.6%)
Black	0 (0.0%)	0 (0.0%)	0 (0.0%)	10,595 (17.6%)
Others	0 (0.0%)	0 (0.0%)	0 (0.0%)	9,515 (15.8%)
<i>Gender</i>				
Female	102,237 (48.1%)	44,124 (48.5%)	39,050 (48.9%)	31,483 (52.2%)
Male	110,481 (51.9%)	46,787 (51.5%)	40,736 (51.1%)	28,799 (47.8%)
<i>Age</i>				
Mean ± standard deviation	59.7 ± 17.7	62.8 ± 18.2	61.3 ± 16.3	58.1 ± 16.7
<50 y/o	100,722 (47.4%)	42,836 (47.1%)	34,585 (43.3%)	18,849 (31.3%)
50–64 y/o	58,737 (27.6%)	25,410 (28.0%)	21,305 (26.7%)	16,429 (27.3%)
≥65 y/o	53,259 (25.0%)	22,665 (24.9%)	23,896 (30.0%)	25,004 (41.5%)
Diabetes, <i>n</i> (%)	11,412 (5.4%)	5,040 (5.5%)	3,480 (4.4%)	7,269 (12.1%)
Hypertension, <i>n</i> (%)	12,101 (5.7%)	5,162 (5.7%)	3,396 (4.3%)	13,639 (22.6%)
Hyperlipidemia, <i>n</i> (%)	9,015 (4.2%)	3,857 (4.2%)	3,536 (4.4%)	8,613 (14.3%)
<i>Number of disease histories^a</i>				
Mean ± standard deviation	7.0 ± 11.3	7.0 ± 11.3	12.2 ± 16.7	7.4 ± 13.3
≤2	104,022 (48.9%)	44,878 (49.4%)	27,748 (34.8%)	35,641 (59.1%)
3–10	61,962 (29.1%)	26,229 (28.9%)	22,611 (28.3%)	9,941 (16.5%)
≥11	46,734 (22.0%)	19,804 (21.8%)	29,427 (36.9%)	14,700 (24.4%)
<i>Number of prevalent phecode</i>				
Mean ± standard deviation	3.2 ± 3.5	3.1 ± 3.5	3.0 ± 3.2	7.4 ± 7.3
≤2	117,777 (55.4%)	50,312 (55.3%)	45,339 (56.8%)	20,671 (34.3%)
3–5	58,405 (27.5%)	24,864 (27.3%)	22,403 (28.1%)	10,266 (17.0%)
≥6	36,536 (17.2%)	15,735 (17.3%)	12,044 (15.1%)	29,345 (48.7%)
<i>Number of incident phecode</i>				
Mean ± standard deviation	5.9 ± 9.1	5.9 ± 9.1	8.7 ± 12.0	6.1 ± 11.3
≤2	109,037 (51.3%)	46,309 (50.9%)	32,425 (40.6%)	36,900 (61.2%)
3–10	63,515 (29.9%)	27,511 (30.3%)	24,767 (31.0%)	11,269 (18.7%)
≥11	40,166 (18.9%)	17,091 (18.8%)	22,594 (28.3%)	12,113 (20.1%)

^aNumber of disease histories refers to the total mapped phecodes recorded prior to or at the index date.

and age, and (d) the *p*-values testing the independent predictive value of the CXR embedding in multivariable models. After adjustment, 554 prevalent and 457 incident phecodes were significantly predicted in all three validation sets, attesting to the broad disease signal in CXR embeddings. Figure 2a shows the AUC distribution for 1074 prevalent classifiers; 60 achieved AUC ≥ 0.85 in hold-out/external and ≥ 0.80 in MIMIC significantly after Bonferroni-correction, led by circulatory (*n* = 21), respiratory (*n* = 12), and neoplasm (*n* = 7) phenotypes. Figure 2b plots C-index for 1074 incident predictors; 42 reached C-index ≥ 0.80 in hold-out/external and ≥ 0.75 in MIMIC, again dominated by circulatory (*n* = 21) and respiratory (*n* = 7) diseases. To further characterize model performance beyond discrimination metrics, we evaluated calibration statistics and predictive values for

representative phenotypes. Calibration performance was summarized using calibration intercept, calibration slope, and Brier score (Supplementary Data 5). As expected for low-prevalence phenotypes, calibration statistics varied across datasets, reflecting differences in case-mix and disease prevalence. In sensitivity analyses accounting for potential confounding from device-related features, model performance remained largely unchanged. Across datasets, discrimination metrics were highly consistent with the primary analysis, with only minimal variations observed in AUC and related measures (Supplementary Data 6 and 7). These findings suggest that the overall predictive performance is robust and not materially driven by device-related or contextual factors, supporting the stability of the learned imaging representations.

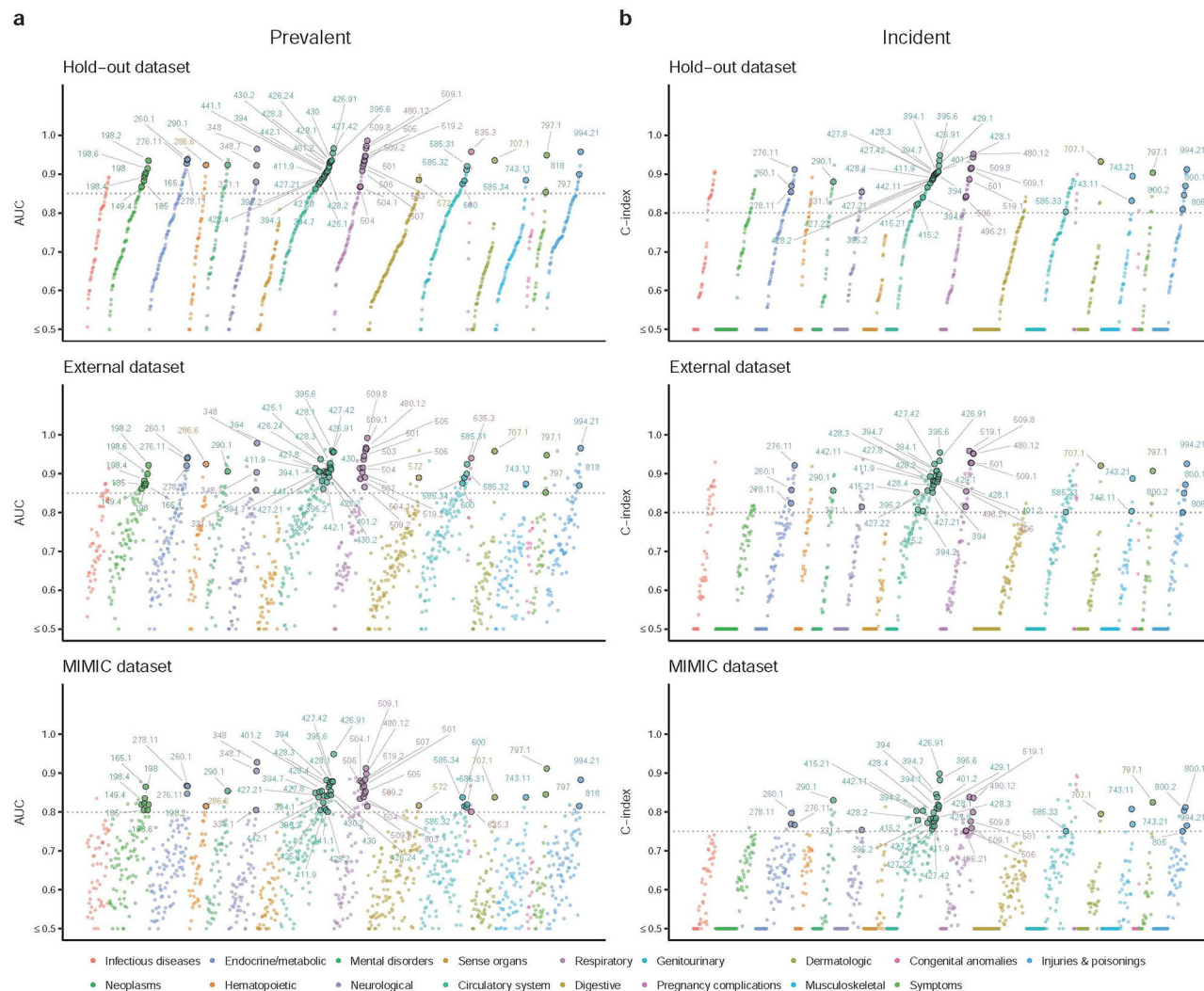


Fig. 2 | Chest X-ray AI model performance on prevalent and incident phecodes across three datasets. a Prevalent disease classifier performance. Each panel shows the distribution of area under the receiver operating characteristic curve (AUC) for classifiers predicting the presence of individual phecodes from chest X-ray. Dots represent individual phecode AUCs, colored by clinical category (see legend) and ordered by AUCs in the hold-out dataset. Horizontal dashed lines mark the performance thresholds (AUC = 0.85 for both hold-out and external dataset; AUC = 0.80 for MIMIC dataset). **b** Incident disease predictor performance. Each panel shows the distribution of the C-index for predictors predicting the future events of

individual phecodes from chest X-ray. Dots represent individual phecode AUCs, colored by clinical category (see legend) and ordered by C-indexes in the hold-out dataset. Horizontal dashed lines mark the performance thresholds (C-index = 0.8 for both hold-out and external dataset; C-index = 0.75 for MIMIC dataset). Only phecodes that (1) achieved performance thresholds in three datasets at the same time, and (2) showed statistically significant improvement (p value < 0.05/1074) after covariate adjustment by imaging type, position, patient race, gender, and age, are highlighted and annotated by their phecode IDs.

Supplementary Fig. 1 reveals clear disparities: using anterior-posterior (AP) vs. PA views caused moderate ($0.05 < \Delta \leq 0.15$) or large ($\Delta > 0.15$) drops in 60% and 15% of tasks, respectively, with only 2%/1% showing improvements. Comparing White vs. Asian patients, 47%/5% of tasks showed moderate/large declines, while 5%/1% improved. Older age groups suffered more drops (35% moderate and 6% large for <50 vs. 50–64; 30% and 1% for 50–64 vs. ≥ 65), with few gains. Patients with fewer disease histories and comorbidities saw better performance than those with more. Acquisition position and gender had minimal impact, and Black/Other races performed comparably to Asian. To provide practical guidance for AP-dominant settings, we performed an AP-specific reliability analysis using the same predefined performance thresholds within AP-view subgroups. In total, 12 prevalent and 17 incident phecodes retained robust AP-view performance across validation cohorts (Supplementary Data 8), including cardiovascular or device/procedure-associated phenotypes such as cardiac pacemaker in situ, heart valve replacement, rheumatic heart valve disease, and tricuspid valve disease. These findings suggest that AP-view

degradation is phenotype-specific and that AP-dominant deployment should be restricted to AP-validated phecodes.

Predictable phenotype grouping

Supplementary Fig. 2 presents the cosine similarity matrix of 512-dimensional regression embeddings for 102 high-accuracy phenotypes (60 prevalent, 42 incident), and hierarchical clustering identified 28 distinct groups. To assess whether these clusters could be explained by real-world disease co-occurrence, we further evaluated pairwise phecode correlations across the EHR using the phi coefficient (Supplementary Fig. 3). Correlations were generally weak and lacked a clear modular structure. This contrasts with the well-defined clusters observed in the embedding space, suggesting that the clustering is not primarily driven by epidemiological co-occurrence, but instead reflects shared radiographic representations captured by the model. Cross-disease transfer within clusters incurs virtually no loss in AUC or C-index (≤ 0.05 drop; Supplementary Fig. 4), indicating shared CXR signatures. Table 2 summarizes their associations:

Table 2 | Top association list

Phocode	Type	Disease	Group	AUC/C-index (hold-out)	AUC/C-index (external)	AUC/C-index (MIMIC)
<i>1. Future cardiovascular disease</i>						
426.91	Incident	Cardiac pacemaker in situ	Circulatory system	0.935 (0.927–0.942)	0.933 (0.926–0.939)	0.898 (0.892–0.905)
427.42	Incident	Cardiac arrest	Circulatory system	0.897 (0.891–0.903)	0.925 (0.919–0.931)	0.764 (0.752–0.775)
797.1	Incident	Cardiogenic shock	Symptoms	0.904 (0.894–0.913)	0.907 (0.895–0.919)	0.825 (0.814–0.836)
395.6	Incident	Heart valve replaced	Circulatory system	0.948 (0.940–0.957)	0.954 (0.946–0.963)	0.882 (0.874–0.890)
426.24	Prevalent	Atrioventricular block, complete	Circulatory system	0.931 (0.908–0.955)	0.915 (0.872–0.958)	0.854 (0.836–0.872)
428.1	Incident	Congestive heart failure (CHF) NOS	Circulatory system	0.907 (0.903–0.911)	0.891 (0.887–0.895)	0.808 (0.802–0.814)
428.3	Incident	Heart failure with reduced EF	Circulatory system	0.901 (0.895–0.907)	0.895 (0.890–0.900)	0.785 (0.778–0.791)
503	Prevalent	Pulmonary congestion and hypostasis	Respiratory	0.926 (0.903–0.948)	0.936 (0.904–0.969)	0.838 (0.825–0.852)
290.1	Incident	Dementias	Mental disorders	0.880 (0.875–0.885)	0.856 (0.852–0.861)	0.830 (0.825–0.836)
429.1	Incident	Heart transplant/surgery	Circulatory system	0.921 (0.901–0.942)	0.897 (0.871–0.923)	0.815 (0.797–0.833)
401.2	Incident	Hypertensive heart and/or renal disease	Circulatory system	0.905 (0.901–0.910)	0.886 (0.881–0.891)	0.819 (0.814–0.823)
428.2	Prevalent	Heart failure NOS	Circulatory system	0.908 (0.889–0.927)	0.904 (0.879–0.929)	0.800 (0.766–0.835)
428.2	Incident	Heart failure NOS	Circulatory system	0.877 (0.861–0.893)	0.893 (0.879–0.907)	0.782 (0.764–0.800)
442.1	Prevalent	Aortic aneurysm	Circulatory system	0.906 (0.883–0.930)	0.878 (0.841–0.914)	0.841 (0.819–0.862)
394	Incident	Rheumatic disease of the heart valves	Circulatory system	0.904 (0.896–0.911)	0.878 (0.870–0.887)	0.838 (0.829–0.847)
394.1	Prevalent	Mitral valve stenosis and aortic valve stenosis	Circulatory system	0.873 (0.840–0.907)	0.906 (0.874–0.938)	0.838 (0.818–0.858)
394.1	Incident	Mitral valve stenosis and aortic valve stenosis	Circulatory system	0.903 (0.890–0.917)	0.907 (0.896–0.919)	0.810 (0.799–0.821)
394.7	Incident	Disease of tricuspid valve	Circulatory system	0.901 (0.888–0.913)	0.906 (0.896–0.916)	0.846 (0.835–0.857)
394.7	Prevalent	Disease of tricuspid valve	Circulatory system	0.877 (0.846–0.909)	0.905 (0.869–0.940)	0.852 (0.836–0.868)
427.8	Prevalent	Sinoatrial node dysfunction (Bradycardia)	Circulatory system	0.862 (0.820–0.904)	0.914 (0.891–0.936)	0.849 (0.831–0.867)
427.8	Incident	Sinoatrial node dysfunction (Bradycardia)	Circulatory system	0.899 (0.887–0.911)	0.880 (0.870–0.891)	0.781 (0.765–0.796)
411.9	Incident	Other acute and subacute forms of ischemic heart disease	Circulatory system	0.896 (0.884–0.908)	0.880 (0.865–0.894)	0.775 (0.764–0.785)
743.21	Incident	Pathologic fracture of vertebrae	Musculoskeletal	0.895 (0.888–0.902)	0.887 (0.881–0.894)	0.769 (0.755–0.783)
427.21	Incident	Atrial fibrillation	Circulatory system	0.886 (0.881–0.891)	0.851 (0.846–0.857)	0.763 (0.758–0.769)
428.4	Prevalent	Heart failure with preserved EF	Circulatory system	0.887 (0.839–0.934)	0.881 (0.846–0.916)	0.854 (0.847–0.860)
428.4	Incident	Heart failure with preserved EF	Circulatory system	0.891 (0.878–0.904)	0.871 (0.858–0.883)	0.799 (0.794–0.804)
395.2	Prevalent	Nonrheumatic aortic valve disorders	Circulatory system	0.890 (0.870–0.910)	0.860 (0.839–0.882)	0.825 (0.813–0.837)
395.2	Incident	Nonrheumatic aortic valve disorders	Circulatory system	0.866 (0.856–0.876)	0.856 (0.848–0.865)	0.772 (0.762–0.782)
442.11	Incident	Abdominal aortic aneurysm	Circulatory system	0.888 (0.870–0.906)	0.882 (0.863–0.901)	0.783 (0.766–0.801)
427.22	Incident	Atrial flutter	Circulatory system	0.885 (0.876–0.894)	0.861 (0.851–0.872)	0.756 (0.746–0.765)
800.1	Incident	Fracture of neck of femur	Injuries & poisonings	0.869 (0.861–0.878)	0.872 (0.864–0.880)	0.812 (0.800–0.824)
800.2	Incident	Fracture of unspecified part of femur	Injuries & poisonings	0.846 (0.833–0.858)	0.850 (0.837–0.862)	0.803 (0.789–0.818)
394.2	Incident	Mitral valve disease	Circulatory system	0.840 (0.817–0.863)	0.804 (0.776–0.831)	0.803 (0.771–0.835)
496.21	Incident	Obstructive chronic bronchitis	Respiratory	0.840 (0.832–0.848)	0.815 (0.808–0.821)	0.751 (0.741–0.760)
415.2	Incident	Chronic pulmonary heart disease	Circulatory system	0.823 (0.810–0.837)	0.807 (0.798–0.817)	0.780 (0.773–0.786)
415.21	Incident	Primary pulmonary hypertension	Circulatory system	0.820 (0.790–0.850)	0.852 (0.832–0.872)	0.803 (0.739–0.867)
805	Incident	Fracture of vertebral column without mention of spinal cord injury	Injuries & poisonings	0.809 (0.801–0.817)	0.800 (0.793–0.807)	0.750 (0.739–0.762)
<i>2. Future critical illness</i>						
509.8	Incident	Dependence on a respirator [Ventilator] or supplemental oxygen	Respiratory	0.943 (0.938–0.948)	0.952 (0.948–0.956)	0.800 (0.791–0.808)
480.12	Prevalent	Pseudomonal pneumonia	Respiratory	0.972 (0.961–0.983)	0.964 (0.948–0.980)	0.898 (0.874–0.922)
480.12	Incident	Pseudomonal pneumonia	Respiratory	0.952 (0.946–0.958)	0.951 (0.945–0.958)	0.836 (0.811–0.862)
509.1	Incident	Respiratory failure	Respiratory	0.916 (0.913–0.919)	0.927 (0.925–0.930)	0.759 (0.753–0.764)
994.21	Incident	Septic shock	Injuries & poisonings	0.912 (0.908–0.916)	0.925 (0.922–0.929)	0.765 (0.757–0.773)
519.2	Prevalent	Respiratory complications	Respiratory	0.944 (0.913–0.975)	0.890 (0.810–0.971)	0.883 (0.872–0.894)
276.11	Incident	Hyperosmolality and/or hyponatremia	Endocrine/metabolic	0.912 (0.903–0.920)	0.922 (0.914–0.929)	0.767 (0.759–0.774)
707.1	Prevalent	Decubitus ulcer	Dermatologic	0.935 (0.921–0.950)	0.958 (0.949–0.966)	0.838 (0.827–0.849)

Table 2 (continued) | Top association list

Phocode	Type	Disease	Group	AUC/C-index (hold-out)	AUC/C-index (external)	AUC/C-index (MIMIC)
707.1	Incident	Decubitus ulcer	Dermatologic	0.932 (0.928–0.937)	0.920 (0.915–0.925)	0.795 (0.787–0.803)
509.2	Prevalent	Respiratory insufficiency	Respiratory	0.935 (0.906–0.964)	0.865 (0.789–0.942)	0.847 (0.826–0.868)
501	Prevalent	Pneumonitis due to inhalation of food or vomitus	Respiratory	0.934 (0.924–0.943)	0.945 (0.932–0.958)	0.867 (0.859–0.874)
501	Incident	Pneumonitis due to inhalation of food or vomitus	Respiratory	0.913 (0.908–0.918)	0.928 (0.924–0.932)	0.776 (0.768–0.784)
506	Incident	Empyema and pneumothorax	Respiratory	0.843 (0.829–0.858)	0.855 (0.840–0.870)	0.752 (0.736–0.769)
441.1	Prevalent	Acute vascular insufficiency of intestine	Circulatory system	0.902 (0.835–0.969)	0.902 (0.823–0.981)	0.805 (0.776–0.834)
519.1	Incident	Tracheostomy complications	Respiratory	0.886 (0.863–0.910)	0.958 (0.949–0.967)	0.838 (0.810–0.867)
331.1	Incident	Hydrocephalus	Neurological	0.854 (0.844–0.864)	0.814 (0.805–0.824)	0.754 (0.730–0.778)
3. Cachexia and advanced malignancies						
260.1	Prevalent	Cachexia	Endocrine/metabolic	0.938 (0.925–0.950)	0.941 (0.925–0.958)	0.867 (0.851–0.882)
260.1	Incident	Cachexia	Endocrine/metabolic	0.870 (0.862–0.878)	0.858 (0.849–0.866)	0.798 (0.788–0.808)
198.2	Prevalent	Secondary malignancy of respiratory organs	Neoplasms	0.934 (0.921–0.948)	0.922 (0.905–0.938)	0.819 (0.803–0.835)
198.6	Prevalent	Secondary malignancy of bone	Neoplasms	0.915 (0.899–0.930)	0.899 (0.873–0.925)	0.805 (0.789–0.822)
198	Prevalent	Secondary malignant neoplasm	Neoplasms	0.898 (0.882–0.915)	0.869 (0.844–0.894)	0.835 (0.823–0.848)
198.4	Prevalent	Secondary malignant neoplasm of liver	Neoplasms	0.892 (0.874–0.910)	0.882 (0.857–0.907)	0.823 (0.808–0.837)
4. Advanced heart failure spectrum						
505	Prevalent	Other pulmonary inflammation or edema	Respiratory	0.947 (0.936–0.958)	0.960 (0.952–0.968)	0.853 (0.838–0.868)
428.1	Prevalent	Congestive heart failure (CHF) NOS	Circulatory system	0.931 (0.925–0.938)	0.926 (0.919–0.934)	0.876 (0.871–0.882)
428.3	Prevalent	Heart failure with reduced EF	Circulatory system	0.928 (0.916–0.940)	0.919 (0.904–0.933)	0.871 (0.864–0.878)
401.2	Prevalent	Hypertensive heart and/or renal disease	Circulatory system	0.915 (0.904–0.927)	0.902 (0.889–0.915)	0.864 (0.848–0.881)
427.21	Prevalent	Atrial fibrillation	Circulatory system	0.893 (0.883–0.903)	0.896 (0.884–0.907)	0.842 (0.837–0.847)
425.1	Prevalent	Primary/intrinsic cardiomyopathies	Circulatory system	0.880 (0.837–0.923)	0.937 (0.909–0.965)	0.805 (0.790–0.820)
5. Hemorrhage and coma						
430	Prevalent	Intracranial hemorrhage	Circulatory system	0.933 (0.922–0.945)	0.911 (0.885–0.936)	0.844 (0.819–0.869)
348.7	Prevalent	Coma	Neurological	0.922 (0.907–0.938)	0.903 (0.877–0.930)	0.906 (0.889–0.923)
430.2	Prevalent	Intracerebral hemorrhage	Circulatory system	0.920 (0.910–0.930)	0.891 (0.871–0.911)	0.846 (0.834–0.858)
818	Prevalent	Intracranial hemorrhage (injury)	Injuries & poisonings	0.899 (0.888–0.911)	0.869 (0.844–0.895)	0.816 (0.799–0.833)
6. Sepsis and infarction						
509.8	Prevalent	Dependence on a respirator [Ventilator] or supplemental oxygen	Respiratory	0.986 (0.981–0.991)	0.992 (0.990–0.994)	0.815 (0.803–0.826)
994.21	Prevalent	Septic shock	Injuries & poisonings	0.957 (0.951–0.963)	0.965 (0.958–0.972)	0.883 (0.876–0.891)
286.6	Prevalent	Defibrination syndrome	Hematopoietic	0.923 (0.894–0.953)	0.924 (0.882–0.967)	0.815 (0.785–0.845)
411.9	Prevalent	Other acute and subacute forms of ischemic heart disease	Circulatory system	0.898 (0.874–0.922)	0.906 (0.869–0.944)	0.810 (0.796–0.824)
7. Cardiac arrest and respiratory failure						
509.1	Prevalent	Respiratory failure	Respiratory	0.966 (0.963–0.969)	0.967 (0.962–0.971)	0.913 (0.908–0.917)
348	Prevalent	Other conditions of brain	Neurological	0.965 (0.950–0.980)	0.979 (0.967–0.990)	0.928 (0.916–0.940)
427.42	Prevalent	Cardiac arrest	Circulatory system	0.954 (0.945–0.962)	0.955 (0.942–0.968)	0.878 (0.867–0.890)
8. Dialysis-dependent disease						
585.31	Prevalent	Renal dialysis	Genitourinary	0.920 (0.908–0.933)	0.924 (0.910–0.938)	0.819 (0.802–0.837)
585.32	Prevalent	End-stage renal disease	Genitourinary	0.911 (0.900–0.921)	0.899 (0.884–0.915)	0.812 (0.796–0.827)
9. Advanced chronic kidney disease						
585.34	Prevalent	Chronic Kidney Disease, Stage IV	Genitourinary	0.883 (0.862–0.903)	0.889 (0.869–0.910)	0.815 (0.799–0.831)
585.33	Incident	Chronic Kidney Disease, Stage III	Genitourinary	0.803 (0.797–0.809)	0.801 (0.796–0.807)	0.751 (0.744–0.757)
10. Rheumatic lung disease						
394	Prevalent	Rheumatic disease of the heart valves	Circulatory system	0.904 (0.885–0.923)	0.904 (0.877–0.931)	0.882 (0.859–0.905)
504	Prevalent	Other alveolar and parietoalveolar pneumonopathy	Respiratory	0.867 (0.811–0.922)	0.914 (0.877–0.951)	0.834 (0.795–0.874)

Table 2 (continued) | Top association list

Phecode	Type	Disease	Group	AUC/C-index (hold-out)	AUC/C-index (external)	AUC/C-index (MIMIC)
<i>11. Osteoporosis</i>						
743.11	Prevalent	Osteoporosis NOS	Musculoskeletal	0.884 (0.871–0.897)	0.874 (0.857–0.890)	0.838 (0.829–0.847)
743.11	Incident	Osteoporosis NOS	Musculoskeletal	0.831 (0.826–0.836)	0.803 (0.799–0.808)	0.808 (0.800–0.815)
<i>12. Prostate disorder</i>						
185	Prevalent	Cancer of prostate	Neoplasms	0.882 (0.868–0.897)	0.873 (0.860–0.886)	0.805 (0.792–0.818)
600	Prevalent	Hyperplasia of prostate	Genitourinary	0.875 (0.868–0.881)	0.875 (0.868–0.883)	0.838 (0.830–0.845)
<i>13. Pacemaker</i>						
426.91	Prevalent	Cardiac pacemaker in situ	Circulatory system	0.966 (0.947–0.986)	0.956 (0.938–0.975)	0.949 (0.942–0.956)
<i>14. Placental complications</i>						
635.3	Prevalent	Placenta previa and abruptio placenta	Pregnancy complications	0.957 (0.910–1.000)	0.940 (0.877–1.000)	0.801 (0.676–0.926)
<i>15. Cardiogenic shock</i>						
797.1	Prevalent	Cardiogenic shock	Symptoms	0.949 (0.931–0.967)	0.948 (0.916–0.979)	0.912 (0.903–0.920)
<i>16. Hypernatremic hyperosmolality</i>						
276.11	Prevalent	Hyperosmolality and/or hypernatremia	Endocrine/metabolic	0.937 (0.921–0.953)	0.939 (0.924–0.954)	0.847 (0.839–0.855)
<i>17. Valve replacement</i>						
395.6	Prevalent	Heart valve replaced	Circulatory system	0.935 (0.884–0.987)	0.959 (0.935–0.984)	0.880 (0.866–0.893)
<i>18. Prevalent morbid obesity</i>						
278.11	Prevalent	Morbid obesity	Endocrine/metabolic	0.927 (0.902–0.953)	0.921 (0.880–0.961)	0.868 (0.860–0.876)
<i>19. Dementias</i>						
290.1	Prevalent	Dementias	Mental disorders	0.924 (0.915–0.932)	0.906 (0.894–0.918)	0.854 (0.848–0.861)
<i>20. Empyema and pneumothorax</i>						
506	Prevalent	Empyema and pneumothorax	Respiratory	0.920 (0.905–0.935)	0.915 (0.901–0.930)	0.863 (0.850–0.876)
<i>21. Pleurisy and effusion</i>						
507	Prevalent	Pleurisy; pleural effusion	Respiratory	0.909 (0.898–0.919)	0.904 (0.891–0.916)	0.876 (0.869–0.883)
<i>22. Bronchial carcinoma</i>						
165.1	Prevalent	Cancer of bronchus; lung	Neoplasms	0.903 (0.892–0.914)	0.874 (0.858–0.890)	0.865 (0.853–0.877)
<i>23. Ascites</i>						
572	Prevalent	Ascites (non-malignant)	Digestive	0.885 (0.864–0.907)	0.890 (0.865–0.914)	0.817 (0.803–0.831)
<i>24. Hydrocephalus</i>						
331.1	Prevalent	Hydrocephalus	Neurological	0.880 (0.852–0.907)	0.858 (0.827–0.889)	0.805 (0.783–0.828)
<i>25. Fibrosing alveolitis</i>						
504.1	Prevalent	Idiopathic fibrosing alveolitis	Respiratory	0.868 (0.826–0.910)	0.887 (0.851–0.922)	0.880 (0.843–0.918)
<i>26. Laryngeal carcinoma</i>						
149.4	Prevalent	Cancer of larynx	Neoplasms	0.867 (0.810–0.924)	0.861 (0.775–0.947)	0.819 (0.766–0.873)
<i>27. Future morbid obesity</i>						
278.11	Incident	Morbid obesity	Endocrine/metabolic	0.854 (0.840–0.868)	0.824 (0.809–0.839)	0.769 (0.762–0.777)
<i>28. Shock</i>						
797	Prevalent	Shock	Symptoms	0.853 (0.847–0.859)	0.851 (0.842–0.861)	0.845 (0.832–0.859)

Displayed are that (1) achieved performance thresholds in three datasets at the same time and (2) showed statistically significant improvement (p -value < 0.05/1074) after covariate adjustment by imaging type, position, patient race, gender, and age.

Cluster 1 (future cardiovascular disease) contains 28 incident phecodes (64% of all incidents; all 21 circulatory codes). Representative tasks in this cluster predict future cardiac event, such as pacemaker (phecode: 426.91; C-index 0.935/0.933/0.898 in hold-out/external/MIMIC), cardiac arrest (427.42; C-index 0.897/0.925/0.764), cardiogenic shock (797.1; C-index 0.904/0.907/0.825), heart valve replacement (395.6; C-index 0.948/0.954/0.882), and incident heart failures (428.1, 428.3, 428.2, and 428.4). The same cluster also contains 9 prevalent phecodes, such as complete atrioventricular block (426.24; AUC 0.931/0.915/0.854) and pulmonary congestion and hypostasis (503; AUC 0.926/0.936/0.838). Unexpectedly, it also forecasts

dementia (290.1; C-index 0.880/0.856/0.830) and fractures (743.21, 800.1, 800.2).

Cluster 2 (future critical illness) spans 10 incident phecodes (6 of 7 respiratory) and 6 prevalent severe-respiratory conditions, including ventilator dependence (509.8; C-index 0.943/0.952/0.800), respiratory failure (509.1; C-index 0.916/0.927/0.759), septic shock (994.21; C-index 0.912/0.925/0.765), pseudomonas pneumonia (480.12; AUC 0.972/0.964/0.898), decubitus (707.1; AUC 0.935/0.958/0.838), and aspiration pneumonitis (501; AUC 0.934/0.945/0.867).

Cluster 3 (cachexia and advanced malignancies) includes cachexia (260.1; prevalent AUC 0.938/0.941/0.867; incident C-index 0.870/0.858/



Fig. 3 | t-SNE projection of task and feature embeddings with disease clusters. Each point represents a learned embedding in two-dimensional space after t-distributed stochastic neighbor embedding (t-SNE) of the multimodal co-embedding vectors for 102 phcode prediction tasks and radiological features. Shape

indicates label type, and color indicates cluster membership. Clusters containing more than one phcode are each assigned a unique color and descriptive name (see list). Clusters containing a single phcode are not given a descriptive name; instead, their phcode ID is annotated adjacent to the point.

0.798) and secondary malignancies of lung (198.2), bone (198.6), liver (198.4), and generic neoplasm (198).

Cluster 4 (advanced heart failure spectrum) covers pulmonary edema (505; AUC 0.947/0.960/0.853), congestive heart failure (428.1; AUC 0.931/0.926/0.876), heart failure with reduced ejection fraction (428.3; AUC 0.928/0.919/0.871), plus hypertensive heart/renal disease (401.2), atrial fibrillation (427.21), and primary cardiomyopathies (425.1).

Cluster 5 (hemorrhage and coma) comprises intracranial hemorrhage (430; AUC 0.933/0.911/0.844), coma (348.7; AUC 0.922/0.903/0.906), intracerebral hemorrhage (430.2), and traumatic intracranial hemorrhage (818).

Clusters 6–12 cover sepsis/infarction, cardiac arrest and respiratory failure, dialysis-dependent disease, advanced chronic kidney disease, rheumatic lung disease, osteoporosis, and prostate disorders, each defined by 2–4 phcodes with AUC ≥ 0.85 (hold-out/external) and ≥ 0.80 (MIMIC), or C-index $\geq 0.80/0.75$.

Clusters 13–28 each contain a single phenotype—prevalent pacemaker, prevalent placental complications, prevalent cardiogenic shock, prevalent hypertensive hyperosmolality, prevalent valve replacement, prevalent morbid obesity, prevalent dementias, prevalent empyema and pneumothorax, prevalent pleurisy and effusion, prevalent bronchial carcinoma, prevalent ascites, prevalent hydrocephalus, prevalent fibrosing alveolitis, prevalent carcinoma, incident morbid obesity, and prevalent shock.

Temporal stability of incident phenotype predictions

Supplementary Fig. 5 demonstrates a clear temporal decline in predictive performance across all 42 incident phcode models. In the 30–90 days window, most models show a positive ΔC -index, indicating better discrimination than their overall C-index; accuracy then falls progressively in

the 90–365, 365–1095, and >1095 days windows. Despite this trend, most models remain within ± 0.10 of their overall C-index at all follow-up intervals. Only a handful of phcodes exhibit declines exceeding 0.10 in ΔC -index at the longest time horizon, indicating that CXR-AI captures imaging signatures associated with near- to intermediate-term clinical manifestation in most diseases.

Explainable analysis

Figure 3 visualizes the shared structure of our multimodal embeddings of all 102 phcodes and 57 radiological features. Tasks and features that belong to the same thematic cluster (colored areas) occupy tight, well-separated islands in this space. Notably, cluster 1 (future cardiovascular disease) co-localizes with radiographic markers of vascular pathology—atherosclerosis, aneurysmal dilatation, pulmonary edema, and prominence of hilar shadows—highlighting their shared embedding signatures. Cluster 1 also sits adjacent to clusters 4 (advanced heart failure spectrum), 8 (dialysis-dependent disease), 9 (advanced chronic kidney disease), and 10 (rheumatic lung disease), whereas cluster 2 (future critical illness) lies closer to clusters 6 (sepsis and infarction) and 7 (cardiac arrest and respiratory failure), reflecting the overlapping imaging features among these critical-illness phenotypes.

Supplementary Fig. 6 presents the cosine similarity matrix of embeddings for radiological features, and hierarchical clustering identified 27 distinct groups. Supplementary Fig. 7 maps the cosine similarity between each radiological feature label and every phcode prediction task in our co-embedding space. Notably, classic fluid-overload signs—pleural effusion, pulmonary edema, and consolidation change—show their highest similarity with future cardiovascular disease (cluster 1) and critical illness (cluster 2). Many other features remain near zero for unrelated tasks, underscoring the



Fig. 4 | Radiological features of pathological changes or anatomical/morphological signs ($n = 34$) contributions to CXR-AI phenotype cluster predictions. The heatmap shows standardized regression coefficients linking radiological features of pathological changes or anatomical/morphological signs (y-axis) to CXR-AI model outputs for each phenotype cluster (x-axis). For clusters 1–12 (which contain multiple diseases), each column value represents the mean coefficient across all

disease-specific models within that cluster. Gray cells denote a coefficient of zero (no detectable association). Across the top three rows are the mean R-square values—measured in the hold-out, external, and MIMIC datasets—indicating the proportion of variance in the AI predictions explained by these radiological features. Color intensity scales from white (~0) to deep red (1), reflecting the relative strength of each feature's contribution to the cluster's prediction.

specificity of our embeddings in linking explicit radiologic findings to their corresponding disease-prediction models.

Figure 4 demonstrates that nearly all CXR-AI phenotype clusters can be explained by 34 recognizable radiological findings of pathological changes or anatomical/morphological signs. These dominant radiological features are consistent with established radiological knowledge, including cardiomegaly and vascular calcification in cardiovascular disease and nodular or ground-glass patterns in malignancy. Clusters 1 (future cardiovascular disease), 4 (advanced heart failure spectrum), 8 (dialysis-dependent disease), 9 (advanced chronic kidney disease), and 10 (rheumatic lung disease) are primarily driven by cardiomegaly (0.26/0.46/0.29/0.20 in clusters 1/4/8/10), atherosclerosis (standardized

regression coefficient 0.23/0.22/0.31 in clusters 1/8/9), and ground glass opacity (0.16/0.32/0.29/0.22/0.27 in clusters 1/4/8/9/10). Cluster 2 (future critical illness), cluster 6 (sepsis and infarction), cluster 7 (cardiac arrest and respiratory failure) are primarily driven by ground glass opacity (0.27/0.45/0.29 in clusters 2/6/7), atelectasis (0.25/0.13 in clusters 2/7), atherosclerosis (0.23 in cluster 2), and emphysematous change (0.16/0.14/0.23 in clusters 2/6/7). Cluster 3 (cachexia and advanced malignancies) is driven by atelectasis (0.27), nodules (0.25), and ground glass opacity (0.24). Cluster 5 (hemorrhage and coma) is driven by hypoinflation (0.42) and emphysematous change (0.23). Cluster 11 (osteoporosis) is driven by atherosclerosis (0.44) and cardiomegaly (0.21). Cluster 12 (prostate disorder) is driven by osteophyte formation (0.47)

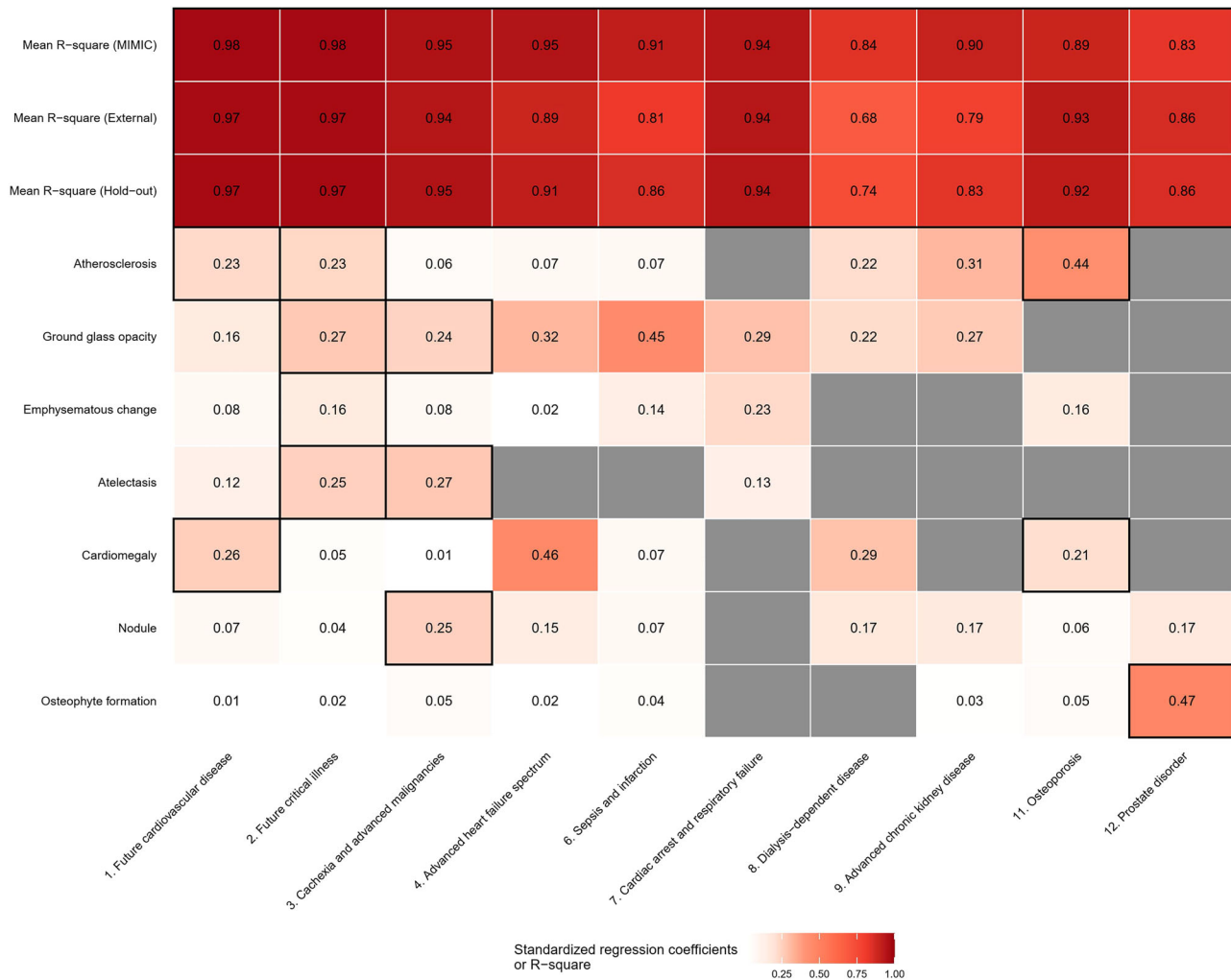


Fig. 5 | Diverse disease phenotypes converge onto a limited set of core radiographic signatures. The heatmap summarizes the dominant radiographic features contributing to each phenotype cluster, highlighting that diverse disease phenotypes map onto a limited set of core imaging signatures. Across clusters, a small number of recurrent features—including atherosclerosis, ground-glass opacity, emphysematous change, atelectasis, and cardiomegaly—consistently explain model predictions. For example, clusters C2, C6, and C7 are primarily characterized by combinations of

atherosclerosis, ground-glass opacity, emphysematous change, and atelectasis, whereas clusters C1, C4, C8, and C9 are dominated by cardiomegaly, atherosclerosis, and ground-glass opacity. Other clusters show similarly compact representations, such as C3 (atelectasis, nodules, and ground-glass opacity), C11 (atherosclerosis and cardiomegaly), and C12 (osteophyte formation and aneurysmal change). These results demonstrate that the apparent diversity of systemic disease phenotypes can be effectively captured by a constrained set of classical radiographic patterns.

and aneurysmal dilatation (0.33). Clusters 13–28 are equally transparent. Supplementary Fig. 8 shows the details of the contribution of radiological features to each disease. To better summarize the relationship between radiographic features and disease phenotypes, we identified the dominant imaging signatures within each phenotype cluster (Fig. 5). Despite the large number of phenotypes evaluated, model predictions were consistently explained by a limited set of recurrent radiographic features, including atherosclerosis, ground-glass opacity, emphysematous change, atelectasis, and cardiomegaly. Multiple phenotype clusters shared highly similar feature compositions. For example, clusters C2, C6, and C7 were primarily characterized by combinations of atherosclerosis, ground-glass opacity, emphysematous change, and atelectasis, whereas clusters C1, C4, C8, and C9 were dominated by cardiomegaly, atherosclerosis, and ground-glass opacity. Other clusters showed similarly compact representations, including C3 (atelectasis, nodules, and ground-glass opacity), C11 (atherosclerosis and cardiomegaly), and C12 (osteophyte formation and aneurysmal change). These findings indicate that the apparent diversity of systemic disease phenotypes can be captured by a constrained set of core radiographic signatures.

Although reconstruction of most CXR-AI phenotype embeddings using radiological features demonstrated high explanatory power (typically $R^2 > 0.85$ across datasets), clusters 8 ($R^2 = 0.74$ – 0.84), 9 ($R^2 = 0.83$ – 0.86), 10 ($R^2 = 0.81$ – 0.90), 14 ($R^2 = 0.63$ – 0.64), 18 ($R^2 = 0.56$), 27 ($R^2 = 0.24$), and 28 ($R^2 = 0.49$ – 0.77) exhibited relatively reduced reconstruction performance, suggesting greater heterogeneity in feature alignment for these phenotypes. Supplemental Table 1 summarizes each phecode's performance after reconstruction: for most tasks, AUC/C-index drops ≤ 0.05 , but one task in cluster 1 (426.91), four in cluster 5 (430.2, 430, 818, 348.7), one in cluster 7 (348), and all tasks in clusters 12, 14, 18, and 27 each decline by >0.05 in at least one dataset. To probe these gaps, we added 23 implant/device and scar features, which moderately improved R^2 (Supplementary Fig. 9). We then refit models with ordinary least squares (OLS) instead of NNLS; this yielded $R^2 \geq 0.98$ (Supplementary Fig. 10) across all tasks but at the cost of allowing negative coefficients. For example, cluster 12 (prostate disorder) is predicted by the absence of prominence of the hilar shadow (standardized regression coefficient -0.32), hypoinflation (-0.27), ICD implantation (-0.25), vertebral body fracture (-0.23), consolidation change (-0.23), and mastectomy (-0.21). This suggests that these clusters may rely on negative inter-disease correlations rather than truly independent radiological predictors.

Discussion

In this work, we applied a contrastively pretrained, multimodal foundation model for CXR and demonstrated structured associations between CXR embeddings and a broad array of clinical phenotypes directly from CXR embeddings. Screening 1074 phecodes, the AI-CXR model significantly predicted 554 prevalent and 457 incident conditions, of which 60 prevalent and 42 incident phenotypes consistently achieved high performance. Hierarchical clustering of the resulting regression-coefficient embeddings revealed 28 coherent phenotype groups, within which cross-phenotype transfer incurs virtually no loss of accuracy. Explainable analyses—leveraging 57 radiologist-curated CXR features—reconstructed most cluster-level AI predictions with high discrimination and pinpointed a small set of signature imaging findings driving each cluster. Together, these results establish a scalable, transferable, and interpretable framework for comprehensive disease profiling from routine chest radiographs.

To predict cardiovascular events from a single CXR may at first seem surprising, but numerous AI-CXR studies have already demonstrated the ability of CXR to forecast left ventricular structural abnormalities²⁷, congestive heart failure²⁸, prevalent/incident low ejection fraction²⁹, and future major adverse cardiovascular events¹⁵. Although many of these investigations focus on cardiovascular outcomes, they often lack interpretability³⁰. Our explainable-AI analysis revealed that the model depends most heavily on markers of cardiomegaly, atherosclerosis, and ground glass opacity, which is highly consistent with previous radiologist observations³¹. These features share a key characteristic: when their classic radiographic patterns are present, they are strongly associated with cardiovascular disease, yet radiologists struggle to perceive their most subtle manifestations^{32,33}. Another advantage of AI lies in quantification—previous work has shown that measuring the extent of aortic arch calcification improves the prediction of new cardiovascular events³⁴. In addition, we observed that although many cardiovascular conditions are predictable from CXR, the underlying models nearly always leverage the same radiological features. In the present study, each phecode was modeled independently using separate linear probes. Future AI-CXR research may explore unified multi-task architectures that explicitly share parameters across related phenotypes rather than training independent disease-specific models. Just as kidney diseases can be predicted by the same set of radiological features. Indeed, prior studies have used CXR to predict future chronic kidney disease and found atherosclerosis to be the most relevant predictor³⁵. Coronary artery diseases and chronic heart failure with volume overload create a pro-arrhythmogenic, fibrotic milieu that underlies atrial fibrillation and atrioventricular block³⁶. Beyond the heart, cardio-renal mechanisms are activated. Congestive heart failure with fluid overload leads to reduced renal perfusion and raised venous pressures, which together precipitate glomerular hypofiltration and renal congestion³⁷. Similarly, rheumatic disease and cardiovascular disease also share common underpinnings involving inflammation³⁸. Most importantly, the explainable AI models achieve high R^2 when reconstructing the original CXR-AI predictions, indicating that in the future, radiologists—once aided by AI to detect those subtle radiological features—could plausibly and accurately explain why these patients carry a high risk of cardiovascular-related diseases. Regression coefficient embeddings reflect clinically calibrated associations derived from real-world labeled data, whereas zero-shot encoding of phecode descriptions would primarily capture linguistic semantic similarity. Direct comparison between these two representational strategies represents an important future direction.

The use of CXR to elucidate the underlying etiologies of acute diseases and symptoms is a time-honored clinical practice, and we found that AI-CXR can accurately predict critical illness events, septic shock, and infarction. Indeed, there have also been studies applying AI-CXR to predict short-term mortality³⁹ and acute respiratory distress syndrome⁴⁰. By applying our explainable analytic approach, we identified that atherosclerosis, ground-glass opacity, emphysematous change, and atelectasis can contribute to future critical illness events. Mechanistically, ground-glass opacities reflect diffuse alveolar injury—whether infectious or edematous—and their severity correlates with progression to respiratory failure⁴¹. Atelectasis,

though often subtle, compromises gas exchange and can precipitate pneumonia and acute respiratory distress syndrome in vulnerable hosts⁴². Aortic arch calcifications on CXR signal systemic atherosclerotic burden and reduced physiological reserve, predisposing patients to future events such as infarction or shock³⁴. In the later fibroproliferative phase of acute respiratory distress syndrome, thoracic computed tomographic scanning often demonstrates features of emphysematous destruction⁴³. Collectively, these findings underscore the dual role of CXR as both a window into chronic comorbidity and a real-time gauge of acute pathology, and they demonstrate that explainable, feature-driven clustering can stratify patients by risk and reveal structured radiological-feature alignment that supports biological plausibility but does not establish mechanistic proof.

CXR is a primary cancer screening tool for bronchial and laryngeal carcinoma, primary respiratory malignancies, secondary bone metastases, and even end-stage cachexia. Our analysis shows that its predictive power stems largely from radiographic signs of atelectasis, nodules, and ground-glass opacities. Pulmonary nodules and ground-glass opacities on chest imaging are classic radiologic indicators of neoplastic processes—often reflecting malignant lesions—and therefore warrant prompt evaluation for advanced malignancy and associated cachexia^{44,45}. Although atelectasis itself is not a cancer-specific finding, in advanced malignancy it often reflects central bronchial obstruction by a tumor—classically producing the “Golden S sign” on chest radiographs—and thus serves as a radiologic marker of highly aggressive disease⁴⁶. Fibrosing alveolitis is another hallmark CXR finding, but radiologists’ accuracy is poor⁴⁷. Likewise, AI can boost pneumothorax detection³. These expected CXR-detectable conditions align with our findings; although AI may enhance diagnostic precision, final model development should be based on confirmatory diagnosis rather than solely on ICD codes.

Osteoporosis is a well-recognized CXR finding, but its detection accuracy without AI support remains relatively low⁴⁸. In our analysis, AI-CXR also leverages associations with age-related conditions—such as atherosclerosis and cardiomegaly⁴⁹—to boost its performance. Indeed, prior saliency-map analyses have shown that AI-CXR models attend to the cardiac region¹⁷, suggesting they may exploit cues outside the ribs to improve accuracy. However, osteoporosis is notoriously underdiagnosed, and a previous randomized controlled trial demonstrated that AI-CXR can meaningfully enhance detection rates¹⁷; therefore, this particular association may still hold valuable clinical utility. In contrast, the link between prostate disorders and CXR is somewhat unexpected: radiologists know that prostate cancer often causes bone metastases, manifesting as osteoblastic changes on CXR⁵⁰, and we observed a weak univariate association in our explainability analysis (Supplementary Fig. 6). However, multivariate analysis revealed that AI-CXR primarily predicts prostate disorders via osteophyte formation and aneurysmal dilatation. Osteophytes typify degenerative joint disease, and aneurysms arise from advanced vascular degeneration. Prostate enlargement often coexists with metabolic and vascular risk factors, further linking these findings⁵¹. This parallels the osteoporosis example: while AI-CXR’s assessment of osteoporosis has been validated in randomized controlled trials, prostate disorders will likely require further study.

In addition to find many diseases can be predicted by CXRs, the key innovation in our study is a transparent, interpretable AI framework built atop the foundation model. Not only does it help radiologists interpret CXRs, but it also exposes spurious disease associations. By using the text encoder’s zero-shot capability, we cataloged an expandable set of common radiological features. In our explainable-AI analysis, these radiological features reconstruct most AI-CXR scores with very high R^2 , indicating radiology has already exhaustively mapped CXR signatures. Future work should map additional diseases to these classic features. In practice, AI can help radiologists locate and quantify these classic features, after which clinicians can integrate them into more accurate, multi-disease predictions. This approach may build trust and resolve the paradox of reduced accuracy in human-AI collaboration^{22,23}.

Our study has several important limitations. First, our interpretability framework is representation-level and descriptive rather than causal. It

characterizes structured associations within the embedding space and evaluates alignment with recognizable radiological features, but does not establish mechanistic causality. ICD-based phecodes reflect documentation timing rather than true biological onset, introducing potential temporal misalignment. Because ICD codes reflect documentation timing rather than true biological onset, it is not possible to fully distinguish subclinical disease present at baseline from true de novo disease development using retrospective EHR-linked data. Although temporal separation and exclusion of prior diagnoses were applied, short-term incident predictions may partially reflect subclinical disease present at baseline imaging. In addition, pixel-level interpretability methods (e.g., Grad-CAM) were not evaluated in this study. While our approach focuses on representation-level alignment between radiological features and disease embeddings, spatial attribution methods may provide complementary insights by localizing image regions contributing to predictions. Future work incorporating such approaches may further enhance clinical interpretability. Second, our explainable AI framework was built on 57 radiologist-curated features. Although these cover many classic CXR findings, they may omit novel imaging biomarkers not typically recorded in conventional radiology reports. Third, multivariable regression analyses may be affected by collinearity among radiological features within the same group, which could lead the model to select only one representative feature and potentially identify statistically driven associations lacking causal relevance. In addition, age is a strong determinant of cardiovascular risk and may contribute to part of the predictive signal captured by CXR embeddings. Furthermore, the stability of feature coefficients across datasets was not formally evaluated and warrants further investigation in future studies. In addition, attribution stability under probe retraining was not formally assessed. Fourth, although we analyzed 1074 phecodes, rare conditions with few cases may yield unstable estimates and wider confidence intervals, potentially overlooking low-prevalence outcomes. Fifth, variability in acquisition protocols and patient populations across institutions may influence feature-based reconstruction stability and overall model performance. In particular, we observed reduced discrimination in AP-view images, which are more prevalent in ICU settings and often reflect greater illness severity. Models deployed in AP-dominant environments may therefore require view-specific recalibration or retraining. Performance derived from predominantly PA-view cohorts should not be directly extrapolated to ICU-heavy or inpatient settings without additional validation. Sixth, demographic composition and potential bias. The internal cohorts were predominantly Asian, whereas the MIMIC cohort included greater racial heterogeneity and consisted entirely of ICU-based imaging. Observed subgroup performance differences may reflect case-mix composition, illness severity, and imaging view distribution (e.g., higher AP prevalence in ICU settings) rather than intrinsic biological variation. These findings should therefore be interpreted as domain shift effects rather than race-specific biological signals. Prospective multi-ethnic validation is required prior to clinical deployment. Finally, while the foundation model demonstrated robust discrimination across 60 prevalent and 42 incident phecodes, its real-world clinical utility remains untested.

In conclusion, we demonstrate that a contrastively pretrained, multimodal foundation model applied to routine chest radiographs can systematically profile a remarkably broad spectrum of human diseases in a scalable, transferable, and interpretable manner. We found that, beyond the traditional uses of CXR, AI-CXR has the potential to predict more than 500 phenotypes, with our most striking discovery being the ability to stratify risk for a broad range of cardiovascular-related diagnoses. Despite profiling over 1000 phenotypes, many conditions converge onto a limited set of classical radiographic signatures within a shared embedding space. This convergence suggests that CXR-derived representations capture fundamental pathophysiologic patterns rather than isolated disease-specific artifacts. Our analytical framework improves representation-level interpretability of AI-CXR by linking phenotype predictions to recognizable radiological feature patterns. However, the present retrospective study does not establish validated patient-level explanations or real-world clinical utility. Future prospective studies should evaluate whether patient-level explanations improve

clinician interpretation, workflow integration, and clinical outcomes. Consequently, we anticipate that future workflows will employ AI to detect and quantify classical CXR findings, allowing radiologists—after review—to interpret a comprehensive disease spectrum through explainable AI. Future prospective studies are warranted to validate its clinical impact.

Methods

Study subjects

The institutional review boards of Tri-Service General Hospital approved this study protocol with a waiver of informed consent (C202305019). This study was conducted in accordance with the Declaration of Helsinki. We performed a retrospective, multi-cohort analysis of de-identified chest radiographs (both posterior–anterior [PA] and anterior–posterior [AP] views) linked to structured clinical data. To avoid patient-level bias, only one radiograph per patient was included—if multiple CXRs were available, a single image was chosen at random.

As shown in Fig. 1a, Tri-Service General Hospital's Neihu General Hospital site (an academic medical center) provided 303,629 patients with at least one CXR during January 2011 to February 2022, which were randomly divided into a development dataset ($n = 212,718$) and a hold-out dataset ($n = 90,911$). To assess generalizability, 79,786 patients from January 2011 to February 2022 were obtained from a geographically distinct community hospital (Tingzhou Branch Hospital) as an external dataset. For further external benchmarking, we used 60,282 CXRs from the MIMIC-CXR database (version 2.0.0) as the MIMIC dataset. For each eligible patient, only one CXR was randomly selected before downstream dataset construction and model training. Thus, each patient contributed a single image to only one cohort, and no images from the same patient could appear across different datasets or cross-validation folds. Patients aged less than 18 years old or without records of the International Classification of Diseases (ICD) were excluded from this study.

Phenotype definitions

ICD-9 and ICD-10 diagnosis codes in each database were mapped to standardized phecodes using the publicly available Phecode library (https://phewascatalog.org/Phecodes_icd10cm)⁵². The timestamp of the CXR examination was used as the index date for merging imaging and clinical records. In the MIMIC-CXR cohort, diagnoses were obtained from the MIMIC-IV (version 2.2) “hosp/diagnoses_icd.csv” and MIMIC-IV-ED (version 2.2) “ed/diagnosis.csv” tables.

In both the hospital EHR and MIMIC systems, ICD codes are recorded at the time a note is completed—either in the outpatient, emergency, or inpatient setting—and therefore may not reflect the exact moment of disease onset. To account for this, we defined: (1) prevalent phecodes as those with a first occurrence within ± 30 days of the index CXR date⁵³. Any patient with a record of the same phecode more than 30 days prior to the index date was excluded from the prevalent disease analysis; (2) Incident phenotypes were defined as the first occurrence of a phecode recorded at least 30 days after the index CXR, thereby excluding diagnoses documented on the same day as imaging. (to avoid overlap with prevalent definitions). Follow-up continued from day 31 post-imaging until the end of available data. Patients with any record of the phecode before the start of incident follow-up were excluded from the incident disease analysis. To ensure robust statistical estimation, we further restricted our analysis to phecodes that had at least five cases (for prevalent phecodes) or five incident events (for incident phecodes) in each of the four datasets (development, hold-out, external, and MIMIC).

Foundation model training and chest radiograph preprocessing

As shown in Fig. 1b, the foundation model was pretrained using a contrastive, multimodal objective on a corpus of 1,004,314 CXR–report pairs⁵⁴. CXRs—comprising both PA and AP views—were resized to 224×224 pixels and encoded by a ViT-B/32 image transformer, while free-text radiology reports were tokenized and encoded by a 12-layer transformer text encoder. Both encoders projected their respective outputs into a shared 512-dimensional embedding space, where positive image–text pairs were pulled

together, and mismatched pairs pushed apart. A softmax-normalized contrastive loss with temperature scaling was used to align radiograph and report embeddings, encouraging semantically meaningful co-embeddings across modalities. At the conclusion of pretraining, the image and text encoders define a transferable foundation model: downstream linear probes (elastic-net regressions) are trained on frozen embeddings, enabling small-sample fine-tuning with as few as 128 positive and 384 negative examples to achieve near–full-dataset performance. Therefore, we used this foundation model to explore extensive phenotype prediction by CXR. Pretraining was conducted for 50 epochs using a fixed training schedule, with validation loss evaluated at the end of each epoch on a held-out 10% subset. The model with the lowest validation loss was selected as the final pretrained model. No dynamic early stopping was applied, as convergence behavior was stable and no evidence of overfitting was observed across training runs. This protocol follows the training procedure described in the original study⁵⁴.

Fine-tuning method on phenotype prediction

We trained a separate linear probe for each phecode using embeddings extracted from the frozen foundation model on the development cohort. Specifically, for each phenotype, we fit an elastic-net-regularized regression on the 512-dimensional image embeddings, tuning the mixing parameter (α) and regularization strength (λ) by five-fold cross-validation to maximize the area under the receiver-operating-characteristic curve (AUC). Training was performed once per phecode, yielding a vector of regression coefficients that capture the relationship between the foundation embeddings and the clinical label. These learned coefficients were then applied without modification to the hold-out, external, and MIMIC datasets for out-of-sample evaluation. No additional fine-tuning or recalibration was performed on any validation dataset, ensuring an unbiased assessment of the foundation model's transferability. Although development dataset images had been included in the pretraining corpus, all samples in the hold-out, external, and MIMIC cohorts were strictly unseen during both pretraining and fine-tuning. The foundation model weights remained frozen during downstream phenotype probe training, minimizing the risk of representation leakage from label supervision.

Statistical analysis

Continuous variables describing patient characteristics in each dataset were summarized as mean $\pm$ standard deviation, and categorical variables as count (percentage). Model discrimination for prevalent phecodes was quantified by the area under the receiver-operating-characteristic curve (AUC)^{55,56}. For incident phecodes, we computed the concordance index (C-index)⁵⁷. To evaluate whether CXR-derived predictions were independently significant, we fitted for each phecode using multivariable logistic regression (prevalent phecodes) or multivariable Cox proportional hazards model (incident phecodes) adjusted by imaging view (PA vs. AP), acquisition position, patient race, gender, and age. Statistical significance was determined using a Bonferroni-corrected threshold of $\alpha = 0.05/1074$ (the total number of tested phecodes).

As illustrated in Fig. 1c, we then identified “high-accuracy” phenotypes meeting all of the following in each of the three validation cohorts (hold-out, external, and MIMIC): (1) prevalent phecodes: AUC ≥ 0.85 (hold-out and external) and ≥ 0.80 (MIMIC); (2) incident phecodes: C-index ≥ 0.80 (hold-out and external) and ≥ 0.75 (MIMIC); and (3) significance at the Bonferroni-corrected level in all three datasets. Discrimination thresholds were selected based on conventional definitions of strong predictive performance in clinical modeling literature and were required to be met consistently across independent validation cohorts with Bonferroni-corrected significance to ensure robustness. Phenotypes satisfying these criteria were carried forward to further analyses.

For these selected phenotypes, we performed stratified assessments across 12 predefined covariates—imaging view, acquisition position, care setting, race/ethnicity, sex, age group, and comorbidity burden. Within each stratum, we compared AUC or C-index between reference and comparison groups, computed the absolute difference (Δ), and reported the proportion

of phenotypes with Δ within ± 0.05 (“acceptable”) or exceeding $\pm 0.05/0.15$ (“moderate”/“large”) bounds.

Furthermore, for incident phecodes, we partitioned follow-up into four windows (30–90, 90–365, 365–1095, and >1095 days) to track CXR-AI C-index over time. In each interval, we excluded patients with total follow-up shorter than the window's lower bound, treated events occurring after the upper bound as censored (non-events), and refit Cox models to derive the interval-specific C-index. For selected representative phenotypes, model calibration was assessed using calibration intercept, calibration slope, and Brier score. Calibration intercept and slope were estimated by regressing observed outcomes on predicted probabilities. The Brier score was used as an overall measure of prediction error. To evaluate the robustness of our findings, we conducted sensitivity analyses. To mitigate potential bias arising from device- or procedure-related signals (e.g., implants, surgical hardware, or indwelling lines), we repeated all primary analyses after excluding individuals with implanted devices or prior surgical interventions.

Explainable analysis

To explore the embedding-level relationships among selected phenotypes and common radiological findings, we first treated each phenotype's vector of elastic-net regression coefficients (512-dimensional) as its representation in the foundation embedding space. Pairwise cosine similarities were computed between all selected phenotype coefficient vectors, and hierarchical clustering was applied to reorganize phenotypes into data-driven groups based on shared imaging predictors. To validate the robustness of each cluster, we applied the model trained on one disease within a given cluster to predict the outcomes of all other diseases in that same cluster, then examined the corresponding changes in AUC or C-index to quantify any performance shifts.

Next, a panel of board-certified radiologists (Wei-Chou Chang, Kai-Hsiung Ko, and Yi-Chih Hsu) curated a list of 57 frequently reported chest radiological features (34 pathological changes or anatomical/morphological signs; 23 implants/devices or surgical scars/traces) using standardized clinical reporting terminology. The 57 radiological features were curated by board-certified radiologists using our institutional chest radiograph reporting terminology, with reference to standard medical and radiology texts, including the Oxford Concise Medical Dictionary and Essentials of Radiology. Synonymous or closely overlapping descriptors were consolidated before feature embedding to ensure consistent representation of radiographic signs. Synonymous descriptors were consolidated into unified feature labels prior to embedding extraction. To derive embeddings for each feature, we fed the foundation model's text encoder two inputs—“<radiological feature>” and “No <radiological feature>”—following the same zero-shot prediction strategy used in prior work⁸. This yields two text embeddings, and we take their difference (“<radiological feature>” minus “No <radiological feature>”) as the feature's embedding. Mathematically, zero-shot prediction compares the dot products of an image embedding with each of these two text embeddings and applies a softmax, which is equivalent to subtracting the text embeddings first and then applying a sigmoid. That sigmoid of the dot product exactly matches the output of a linear probe (logistic regression) using the difference vector as a coefficient, so our computed feature embeddings are rigorously identical to the regression weights one would obtain by linear probing—and can thus be analyzed in the same embedding space.

For joint visualization, we concatenated the phenotype coefficient embeddings and radiological feature embeddings and applied t-distributed stochastic neighbor embedding (t-SNE) to project them into two dimensions. Radiological feature embeddings were derived from the pretrained text encoder using standardized terminology and were independent of site-specific manual annotation. We also calculated cosine similarities between every radiological feature and every selected phenotype embedding to quantify their pairwise associations. Phenotype clustering was performed using cosine similarity of regression coefficient embeddings and was not derived from EHR co-occurrence frequencies. By mapping phenotypes and radiological findings in the same low-dimensional space and examining

their similarity scores, we gain interpretable insights into which specific image features the model leverages when predicting each disease.

We also construct an “explainable AI mode” by combining radiological feature embeddings. Specifically, we first take the 512-dimensional image embeddings from the development dataset and compute their dot product with each radiological feature embedding (which, as noted earlier, are mathematically equivalent to the linear-probe regression coefficients), thereby producing a “radiological feature score” for each CXR. We then use these feature scores to predict the CXR-AI output for each phecode, fitting the model with non-negative least squares (NNLS) to avoid negative coefficients (which are difficult to interpret). After fitting, we extract the regression coefficient for each radiological feature on each phecode and treat the standardized coefficients as measures of independent association. Finally, we validate the explanatory power (R^2) of these models and quantify any changes in AUC or C-index caused by using the “explainable AI model” on the hold-out, external, and MIMIC datasets. In the primary analysis, we used only 34 radiological features representing pathological changes or anatomical/morphological signs, but we also report results from an expanded set that includes an additional 23 features for implants/devices and surgical scars/traces. We also present results obtained using ordinary least squares (OLS) fitting for comparison.

Data availability

Individual-level patient data are not publicly released for privacy reasons. Researchers may request de-identified tabular datasets from the corresponding author after obtaining approval from the Tri-Service General Hospital Institutional Review Board (IRB). Applications should be submitted via the hospital’s Clinical Trial Management System (<https://tsgh.cims.tw/wiPtms/index.html>); the IRB review usually takes about 2–3 months. Once approval is granted, investigators receive VPN credentials that permit secure, on-site analysis. Only aggregated summary results may be exported, and all safeguards required by the IRB remain in place to ensure confidentiality and prevent re-identification. The pretrained foundation model weights are not publicly released due to licensing constraints. However, access to model weights may be granted upon reasonable request under appropriate data use and collaboration agreements, in accordance with institutional policies and current practices for large-scale medical foundation models.

Code availability

The full analysis pipeline, including data preprocessing, feature extraction, and downstream modeling code, is publicly available at: https://github.com/00800217/CXR_PheWAS. This repository provides sufficient detail to reproduce all reported analyses, including phenotype definition, embedding-based modeling, and evaluation procedures. In addition, an archived version of the same codebase is available at <https://doi.org/10.5281/zenodo.19244582>, which provides a persistent and citable snapshot of the repository to ensure long-term accessibility and reproducibility. While pretrained model weights are not publicly distributed, the provided code enables full computational reproducibility given appropriate access to the model and data.

Received: 21 November 2025; Accepted: 2 July 2026;

Published online: 16 July 2026

References

- Smith-Bindman, R. et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996–2010. *JAMA* **307**, 2400–2409 (2012).
- Raoof, S. et al. Interpretation of plain chest roentgenogram. *Chest* **141**, 545–558 (2012).
- Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).
- Willemink, M. J. et al. Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
- Joskowicz, L., Cohen, D., Caplan, N. & Sosna, J. Inter-observer variability of manual contour delineation of structures in CT. *Eur. Radiol.* **29**, 1391–1399 (2019).
- Lutnick, B. et al. An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nat. Mach. Intell.* **1**, 112–119 (2019).
- Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
- Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
- Huang, W. et al. Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning. *Nat. Commun.* **15**, 7620 (2024).
- Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* **14**, 4542 (2023).
- Weiss, J. et al. Deep learning to estimate lung disease mortality from chest radiographs. *Nat. Commun.* **14**, 2797 (2023).
- Chen, Y. C. et al. Integrating VAI-assisted quantified CXRs and multimodal data to assess the risk of mortality. *J. Imaging Inform. Med.* **38**, 1581–1593 (2025).
- Mitsuyama, Y. et al. Chest radiography as a biomarker of ageing: artificial intelligence-based, multi-institutional model development and validation in Japan. *Lancet Healthy Longev.* **4**, e478–e486 (2023).
- Liao, H. C., Lin, C., Wang, C. H. & Fang, W. H. The deep learning algorithm estimates chest radiograph-based sex and age as independent risk factors for future cardiovascular outcomes. *Digit. Health* **9**, 20552076231191055 (2023).
- Weiss, J. et al. Deep learning to estimate cardiovascular risk from chest radiographs: a Risk Prediction Study. *Ann. Intern. Med.* **177**, 409–417 (2024).
- Ma, D., Pang, J., Gotway, M. B. & Liang, J. A fully open AI foundation model applied to chest radiography. *Nature* **643**, 488–498 (2025).
- Lin, C. et al. Osteoporotic precise screening using chest radiography and artificial neural network: the OPSCAN Randomized Controlled Trial. *Radiology* **311**, e231937 (2024).
- Nam, J. G. et al. AI improves nodule detection on chest radiographs in a health screening population: a Randomized Controlled Trial. *Radiology* **307**, e221894 (2023).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- Baselli, G., Codari, M. & Sardanelli, F. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? *Eur. Radiol. Exp.* **4**, 30 (2020).
- Wang, F., Kaushal, R. & Khullar, D. Should health care demand interpretable artificial intelligence or accept “Black Box” medicine? *Ann Intern Med* **172**, 59–60 (2020).
- Seah, J. C. Y. et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit. Health* **3**, e496–e506 (2021).
- Goh, E. et al. Large language model influence on diagnostic reasoning: a Randomized Clinical Trial. *JAMA Netw. Open* **7**, e2440969 (2024).
- Saporta, A. et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat. Mach. Intell.* **4**, 867–878 (2022).
- Çalli, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G. & Murphy, K. Deep learning for chest X-ray analysis: a survey. *Med. Image Anal.* **72**, 102125 (2021).
- Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M. W. & Wiens, J. Deep learning applied to chest X-rays: exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pp. 750–782 (PMLR, 2020).

27. Bhawe, S. et al. Deep learning to detect left ventricular structural abnormalities in chest X-rays. *Eur. Heart J.* **45**, 2002–2012 (2024).
28. Seah, J. C. Y., Tang, J. S. N., Kitchen, A., Gaillard, F. & Dixon, A. F. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology* **290**, 514–522 (2019).
29. Hsiang, C. W. et al. Detection of left ventricular systolic dysfunction using an artificial intelligence-enabled chest X-ray. *Can. J. Cardiol.* **38**, 763–773 (2022).
30. Lauzier, P. T. & Chow, B. J. W. Artificial intelligence detection of left ventricular systolic dysfunction using chest X-rays: prospective validation, please. *Can. J. Cardiol.* **38**, 720–722 (2022).
31. Chait, A., Cohen, H. E., Meltzer, L. E. & VanDurme, J. P. The bedside chest radiograph in the evaluation of incipient heart failure. *Radiology* **105**, 563–566 (1972).
32. Messerli, M. et al. Diagnostic accuracy of chest X-ray dose-equivalent CT for assessing calcified atherosclerotic burden of the thoracic aorta. *Br. J. Radiol.* **90**, 20170469 (2017).
33. Wong, H. Y. F. et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. *Radiology* **296**, E72–e78 (2020).
34. Iijima, K. et al. Aortic arch calcification detectable on chest X-ray is a strong independent predictor of cardiovascular events beyond traditional risk factors. *Atherosclerosis* **210**, 137–144 (2010).
35. Chen, K. C. et al. Prediction of future risk of moderate to severe kidney function loss using a deep learning model-enabled chest radiography. *J. Imaging Inform. Med.* <https://doi.org/10.1007/s10278-025-01489-4> (2025).
36. Nattel, S. Molecular and cellular mechanisms of atrial fibrillation. *JACC Clin. Electrophysiol.* **3**, 425–435 (2017).
37. Mullens, W. et al. Importance of venous congestion for worsening of renal function in advanced decompensated heart failure. *J. Am. Coll. Cardiol.* **53**, 589–596 (2009).
38. Crowson, C. S. et al. Rheumatoid arthritis and cardiovascular disease. *Am. Heart J.* **166**, 622–628.e621 (2013).
39. Chen, Y. J. et al. An AI-enabled dynamic risk stratification for emergency department patients with ECG and CXR integration. *J. Med. Syst.* **47**, 81 (2023).
40. Pai, K. C. et al. Artificial intelligence-aided diagnosis model for acute respiratory distress syndrome combining clinical data and chest radiographs. *Digit. Health* **8**, 20552076221120317 (2022).
41. Cozzi, D. et al. Ground-glass opacity (GGO): a review of the differential diagnosis in the era of COVID-19. *Jpn. J. Radiol.* **39**, 721–732 (2021).
42. Matthay, M. A. et al. Acute respiratory distress syndrome. *Nat. Rev. Dis. Primers* **5**, 18 (2019).
43. Treggiari, M. M., Romand, J. A., Martin, J. B. & Suter, P. M. Air cysts and bronchiectasis prevail in nondependent areas in severe acute respiratory distress syndrome: a computed tomographic study of ventilator-associated changes. *Crit. Care Med.* **30**, 1747–1752 (2002).
44. Seo, J. B., Im, J. G., Goo, J. M., Chung, M. J. & Kim, M. Y. Atypical pulmonary metastases: spectrum of radiologic findings. *Radiographics* **21**, 403–417 (2001).
45. Goo, J. M., Park, C. M. & Lee, H. J. Ground-glass nodules on chest CT as imaging biomarkers in the management of lung adenocarcinoma. *AJR Am. J. Roentgenol.* **196**, 533–543 (2011).
46. Gupta, P. The Golden S sign. *Radiology* **233**, 790–791 (2004).
47. Tung, K. T. et al. Accuracy of the typical computed tomographic appearances of fibrosing alveolitis. *Thorax* **48**, 334–338 (1993).
48. Tsai, D. J. et al. Artificial intelligence-enabled chest X-ray classifies osteoporosis and identifies mortality risk. *J. Med. Syst.* **48**, 12 (2024).
49. Lange, V. et al. The association between bone quality and atherosclerosis: results from two large population-based studies. *Int. J. Endocrinol.* **2017**, 3946569 (2017).
50. Lin, S. C., Yu-Lee, L. Y. & Lin, S. H. Osteoblastic factors in prostate cancer bone metastasis. *Curr. Osteoporos. Rep.* **16**, 642–647 (2018).
51. Suzuki, Y. et al. Benign prostatic hyperplasia and incident cardiovascular disease. *Circ. J.* **88**, 408–416 (2024).
52. Denny, J. C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
53. Choi, Y. et al. Acute and post-acute respiratory complications of SARS-CoV-2 infection: population-based cohort study in South Korea and Japan. *Nat. Commun.* **15**, 4499 (2024).
54. Chen, K. C. et al. A Pretraining Approach for Small-sample Training Employing Radiographs (PASTER): a multimodal transformer trained by chest radiography and free-text reports. *J. Med. Syst.* **49**, 120 (2025).
55. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**, 1315–1316 (2010).
56. Çorbacıoğlu, ŞK. & Aksel, G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: a guide to interpreting the area under the curve value. *Turk. J. Emerg. Med.* **23**, 195–198 (2023).
57. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clin. Kidney J.* **14**, 49–58 (2021).

Acknowledgements

We thank the radiology technologists, clinical informatics teams, and administrative staff of Tri-Service General Hospital for their assistance in data management and infrastructure support throughout this project. The authors also appreciate the constructive feedback from colleagues during internal presentations of this work.

Author contributions

C.L. and K.-C.C. developed the foundation model, performed analyses, and contributed equally to the methodological design. J.-W.H., W.-H.F., W.-C.C., K.-H.K. and Y.-C.H. curated clinical and radiological data and provided expert interpretation. C.-S.L. and S.-H.L. contributed to clinical insight and project oversight. D.-J.T. conceived and supervised the study, coordinated the research activities, and led manuscript preparation. All authors reviewed and approved the final version of the manuscript.

Funding

This study was supported by funding from the National Science and Technology Council, Taiwan (NSTC 112-2222-E-016-001-MY2 to D.J. Tsai, NSTC 114-2321-B-016-005 and NSTC 115-2321-B-016-003 to S.H. Lin).

Competing interests

Author Chin Lin is Associate Editor of npj Digital Medicine. Chin Lin was not involved in the journal's review of, or decisions related to, this manuscript. The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-026-02990-y>.

Correspondence and requests for materials should be addressed to Dung-Jang Tsai.

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) 2026