A Pan-Organ Vision-Language Model for Generalizable 3D CT Representations

Cameron Beeche, BS[1,2]; Joonghyun Kim[1]; Hamed Tavolinejad, MD[2,3]; Bingxin Zhao, PhD[4]; Rakesh Sharma, BS[1]; Jeffrey Duda, PhD[5]; James Gee, PhD[5]; Farouk Dako, MD[5]; Anurag Verma, PhD[6,7]; Colleen Morse MS[6,7]; Bojian Hou, PhD[8]; Li Shen, PhD[8]; Hersh Sagreiya, MD[5]; Christos Davatzikos, PhD[9]; Scott Damrauer, MD[10,11,12]; Marylyn D. Ritchie, PhD[10]; Daniel Rader, MD[6,7,10]; Qi Long, PhD[8]; Tianlong Chen, PhD[13]; Charles E. Kahn, Jr., MD, MS[5]; Julio Chirinos, MD, PhD[2,3]; Walter R. Witschey, PhD[5*]; Penn Medicine Biobank †

[1]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, 19104, USA.

[2]Division of Cardiovascular Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA, 19104, USA.

[3]University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, Philadelphia, PA, 19104, USA.

[4]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, 19104, USA.

[5]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

[6]Department of Medicine, Division of Translational Medicine and Human Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA.

[7]Institute for Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA.

[8]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.

[9]AI2D Center for AI and Data Science, University of Pennsylvania, Philadelphia, USA

[10]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.

[11]Department of Surgery, Division of Vascular Surgery and Endovascular Therapy, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.

[12]Department of Surgery, Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA 19104, USA.
[13]Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

† A full list of contributions from Penn Medicine BioBank team is provided in supplement

***Corresponding author:**
Walter R. Witschey, PhD
South Tower, Rm. 11-155.
Perelman Center for Advanced Medicine.
3400 Civic Center Blvd. Philadelphia, PA. 19104.
Email: witschey@pennmedicine.upenn.edu

Abstract

Generalizable foundation models for computed tomographic (CT) medical imaging data are emerging AI tools anticipated to vastly improve clinical workflow efficiency. However, existing models are typically trained within narrow imaging contexts, including limited anatomical coverage, contrast settings, and clinical indications. These constraints reduce their ability to generalize across the broad spectrum of real-world presentations encountered in volumetric CT imaging data. We introduce Percival, a vision-language foundation model trained on over 400,000 CT volumes and paired radiology reports from more than 50,000 participants enrolled in the Penn Medicine BioBank. Percival employs a dual-encoder architecture with a transformer-based image encoder and a BERT-style language encoder, aligned via symmetric contrastive learning. Percival was validated on over 20,000 participants imaging data encompassing over 100,000 CT volumes. In image-text recall tasks, Percival outperforms models trained on limited anatomical windows. To assess Percival's clinical knowledge, we evaluated the biologic, phenotypic and prognostic relevance using laboratory-wide, phenome-wide association studies and survival analyses, uncovering a rich latent structure aligned with physiological measurements and disease phenotypes.

Keywords: Computed tomography, foundation model, vision-language model, Penn Medicine BioBank.

INTRODUCTION

Computed tomography (CT) is a widely used clinical tool for diagnosis, screening, and monitoring across organ systems and clinical indications, including trauma, cancer, and cardiovascular disease[1-3]. Advances in imaging technology, broader clinical applications, an aging population, and rising disease burden have led to a marked increase in CT imaging examinations[4]. The growing demand for CT imaging, combined with a relatively stagnant growth in the number of radiologists and limited access in rural and resource-constrained settings, has placed pressure on clinical workflows, necessitating more efficient approaches to radiological interpretation and clinical decision support tools[5-7].

Artificial intelligence (AI) can enhance radiological workflows through tools such as automated disease detection, report pre-drafting, and radiomics-guided precision medicine[8-11]. However, clinical deployment of AI remains limited, as task-specific models often fail to generalize beyond narrow applications, highlighting the need for more adaptable solutions[12]. Vision-language models (VLMs) have emerged as a promising framework for learning generalizable representations from large-scale paired image-text data[13-16]. By aligning three-dimensional CT volumes with radiology reports in a latent feature space using contrastive learning, VLMs offer a scalable path toward generalist medical AI. Yet training such models on high-dimensional volumetric data presents unique challenges, including the need for large, diverse datasets that capture the full range of clinical variability. Recent work has made important strides in the feasibility of CT-based VLMs[17,18]; however, further investigation is needed to understand their utility in downstream clinical settings, particularly an investigation of VLM's learned feature representation capabilities for biological alignment, disease label prediction and risk stratification.

Leveraging the Penn Medicine BioBank (PMBB), we developed Percival, a foundation model for three-dimensional CT imaging data. Percival was trained on more than 400,000 three-dimensional CT volumes paired with corresponding radiology reports featuring a wide range of anatomical regions, including the thorax, abdomen, pelvis, head/neck, brain, and extremities, across varying anatomical regions (multi-view) and images with and without iodinated contrast (multi-contrast). In this study, we

investigated four major questions. First, we considered domain generalization by determining the value of multi-view, multi-contrast pretraining compared to view-specific pretraining on six recall tasks. Second, we evaluated the extent to which Percival's generalizable, latent feature space aligns with clinical and biological measurements at the time of imaging. Third, we developed disease phenotype classification models that exploit Percival's latent feature space with no additional pretraining, and we evaluated its out-of-sample performance on an external dataset. Fourth, we explored the prognostic value of Percival's latent feature space, and we identified disease-specific risk stratification profiles. An overview of the study design is presented in **Figure 1**. To support continued research, we have publicly released Percival's pretrained weights, classification and prognostic models: (https://github.com/cams2b/percival).

RESULTS

The study cohort included 73,115 PMBB participants split into a training set (n=51,240) and a validation set (n=21,875). The training and validation cohorts had mean ages of 57±15.9 and 54.9±16.2, with balanced sex distributions. Both cohorts were racially diverse, with approximately 30% and 25% non-white participants, respectively. Patient demographics are provided in **Table S1**. Percival was trained on 402,958 three-dimensional CT volumes, with the majority acquired from Siemens (81.6%) and GE HealthCare (16.9%) scanners. The median slice thickness across CT volumes was 3 millimeters (75% IQR=(1.25mm, 5 mm)). The most frequently encountered CT volume was contrast-enhanced thoracic CT (number of volumes=103,566). See **Tables S2–S3** for summaries of training and validation imaging data.

To interrogate the learning topology of Percival's latent feature space and to evaluate its capacity for unsupervised differentiation among CT volume types and patient demographics, we performed dimensionality reduction using UMAP on a validation set of CT volumes (number of volumes=55,398). UMAP visualizations revealed distinct clusters by CT FOV, indicating that Percival captures modality-specific features (**Fig. S1A**). We also observed distinct sex-specific clusters, indicating that Percival's latent space reflects biological sex differences (**Fig. S1B**). Within specific FOVs, we observed clear sex-specific clustering, most prominently in abdominal and

pelvic CT volumes (**Figs. S1C–S1F**). Additionally, age-specific clustering patterns were evident across FOV and contrast protocols (**Figs. S1G–S1J**).

Recall performance

Image-text recall assesses whether a VLM's latent space captures meaningful cross-modal associations. Strong zero-shot performance indicates clinically relevant features with broad downstream utility. To quantify the value of multi-view, multi-contrast pretraining, we compared Percival with two equivalent FOV-specific models: FOV-Thorax (95,043 non-contrast thoracic CTs; n=14,371) and FOV-AbdPel (77,290 contrast-enhanced abdominal/pelvic CTs; n=20,993). To determine domain generalizability, recall performance was assessed across six distinct tasks spanning two FOVs (thorax and abdomen/pelvis), each evaluated under three contrast modalities: non-contrast, contrast-enhanced, and CT angiogram (**Fig. 2**; **Supplemental Table 4**). Across all tasks, Percival achieved equivalent or superior zero-shot recall performance compared to the FOV-specific baselines (**Figs. 2A-2C**). On the non-contrast thoracic CT recall task, Percival achieved 50.2% top-5 recall accuracy (rank=18.0±34.7) versus 42.8% for FOV-Thorax (rank=25.0±41.9). On contrast-enhanced abdominal/pelvic CT recall task, Percival achieved 43.9% top-5 recall accuracy (rank=17.1±29.5) versus 42.9% for FOV-AbdPel (rank=19.4±33.1).

We next evaluated Percival against Merlin[17], a CT-VLM using a 3D-ResNet image encoder trained on abdominal CT volumes. Percival outperformed Merlin across all tasks, with improvements on abdominal and pelvic tasks (**Figs. 2D–2F**). Specifically, in 256-way recall on contrast-enhanced abdominal and pelvic CT volumes, Percival achieved a top-5 accuracy of 43.9% (rank=17.1±29.5), compared to 35.6% (rank=29.7±43.9) for Merlin. Task-specific rank results are reported for each architecture in **Fig. 2G**.

Biological associations with Percival's latent space

Percival's latent feature embeddings were reduced using principal components (PCs) with the first ten PCs explaining over 50% of the latent feature embedding variance (**Fig. S2**). PCs were significantly associated with age (**Table S5**), with PC1 showing the

strongest relationship (R^2=0.32; **Fig. 3A**). Using a half-half discovery-replication design, we observed significant associations between Percival's latent PCs and seven clinical measurements (**Table S6**) with all significant findings being concordant and in the same direction (**Fig. S3**). Notably, weight and pulse were associated with the greatest number of PCs (8 and 7 PCs, respectively), followed by height and oxygen saturation (4 PCs each), respiratory rate (3 PCs), and temperature (2 PCs).

We then conducted a laboratory-wide association study (LabWAS) using a half-half discovery-replication design to uncover the biological associations between principal components (PCs 1–10) derived from Percival's latent feature embeddings (**Fig. 3B**; **Table S7**). Following FDR correction at the 5% threshold, our discovery cohort (average n=2,183) identified 155 significant associations with 43 laboratory measurements, with 103 associations being further verified in the replication cohort (average n=2,118). Moreover, all associations that passed significance among both the discovery-replication cohorts were concordant and in the same direction (**Fig. S4**). Percival's latent feature space demonstrated consistent associations with laboratory markers of renal function, including blood urea nitrogen (BUN), creatinine, and estimated glomerular filtration rate (eGFR)**.** These markers reflect underlying glomerular and tubular function and are frequently used to assess kidney health[19-21]. Several latent PCs showed significant associations with liver-related and biliary lab measurements, including alkaline phosphatase (ALP), total bilirubin, and urobilinogen. These markers are primary indicators of cholestasis and bile metabolism. Elevated levels often lead to follow-up CT imaging to evaluate for biliary obstruction or hepatic dysfunction[22,23]. There were several associations with glucose, HbA1c, total cholesterol, HDL, LDL, triglycerides, and iron. These markers reflect metabolic health, glycemic control, and lipid regulation. Finally, there were significant associations observed for C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), prothrombin time (PT), and partial thromboplastin time (PTT). CRP and ESR are established biomarkers of systemic inflammation and have been consistently linked to cardiovascular disease risk and atherosclerosis[24,25].

Phenome-wide disease enrichment patterns and classification

To identify downstream targets for Percival's fine-tuning, we performed a PheWAS across 641 ICD-9CM phecodes using 10 latent PCs from the PMBB validation cohort (average n=20,912), revealing numerous disease enrichment patterns (**Fig. 3C, Table S8**). We identified 1,861 significant associations across 521 disease states following FDR correction at the 5% threshold, including 116 significant associations with circulatory system disease states, 65 significant associations with respiratory diseases, 60 associations with neoplasms, and 74 associations with endocrine/metabolic diseases.

We next developed disease classification models for 307 conditions using Percival's latent PCs, adjusting for age and sex (**Fig. 4A**). Specifically, we (**1**) evaluated model performance using five-fold cross-validated AUROC, and (**2**) quantified the added predictive value of Percival beyond age and sex using the DeLong test[26]. In total, 268 classification models exhibited significant AUROC performance gains following Bonferroni correction ($P<0.05/307$; **Table S9**). Specifically, 63 circulatory disease models (**Fig. 4C**) demonstrated performance gains with Percival's integration (AUC mean=0.76; AUC range=(0.67, 0.86)). Notably, Percival's prediction of heart failure with reduced ejection fraction (HFrEF) exhibited an AUC of 0.81 (95% CI=0.79, 0.84). Among vascular diseases, Percival's coronary atherosclerosis and aortic aneurysm models exhibited five-fold AUCs of 0.80 (95% CI=0.79, 0.82) and 0.80 (95% CI=0.76, 0.85) respectively. Interestingly, Percival demonstrated strong performance in predicting atrial fibrillation, with an AUC of 0.79 (95% CI=0.77, 0.80). Respiratory diseases showed strong performance, achieving significant results across 39 conditions (**Fig. 4B**) with a mean AUC of 0.73 (range:0.62, 0.86). Specifically, Percival's classification of pleural effusion exhibited an AUC of 0.79 (95% CI=(0.77, 0.80)), whereas classification of pneumonia had an AUC of 0.74 (95% CI=0.72, 0.75). Interestingly, Percival exhibited markedly strong performance when predicting respiratory failure (AUC=0.84; 95% CI=(0.83, 0.86)).

In brief, Percival's prediction of neoplasms exhibited slightly reduced overall performance compared to other disease categories (AUC mean=0.71; AUC range=(0.61, 0.83)); however, 24 models exhibited performance gains following Percival's integration when compared to the baseline model (**Fig. 4D**). Notable

performance was observed for prediction of benign neoplasms in soft/connective tissues (AUC=0.83; 95% CI=0.78, 0.87), as well as malignant lymph nodes (AUC=0.79; 95% CI=(0.75, 0.82)). We further identified 55, 35, 40 and 12 significant performance gains among classification models for digestive, genitourinary, endocrine/metabolic and musculoskeletal diseases respectively (**Figs. 4E-4G**). Notably, several non-standard CT indications showed strong classification, including morbid obesity (AUC=0.90), diabetes mellitus (AUC=0.75), and cachexia (AUC=0.88). We evaluated Percival's classification models on the CT-RATE dataset using four overlapping diseases (**Fig. 4H**). External AUCs were comparable or better, with pleural effusion (AUC=0.84) and cardiomegaly (AUC=0.83) performing best.

Prognostic modeling

Using survival analysis, we identified prognostic risk signatures across the phenome in which Percival exhibited significant ability to identify elevated and reduced risk profiles for the development of future disease states (**Fig. 5**; **Table S10**). We examined the prognostic utility of Percival's latent PCs in a twofold approach. First, we developed Cox proportional-hazards models for 678 disease states and assessed model performance via five-fold cross-validated concordance-index (**Fig. 5A-5F**). Next, we quantified the utility of the linear predictor to develop disease-specific risk stratification profiles identifying low- intermediate- and high-risk strata (**Fig. 5G**). Risk stratification was measured using the log-rank test Bonferroni adjusted for multiple comparisons (P<0.05/678). Among respiratory diseases (**Fig. 5A**), Percival demonstrated significant risk stratification across 73 disease states with an overall mean C-index of 0.67 (range=(0.51, 0.87)). Notably, Percival predicted future bacterial pneumonia with a C-index of 0.77 (95% CI=(0.76, 0.78); disease enrichment=4.12).

When examining circulatory system disease states, (**Fig. 5B**) Percival exhibited significant risk stratification across 126 disease states (C-index mean=0.70; C-index range=(0.56, 0.87)). Interestingly, we observed significant performance and risk stratification for predicting future heart failure with preserved ejection fraction (HFpEF) (C-index=0.78; 95% CI=(0.77, 0.79); enrichment=5.77). When examining vascular disease, Percival was predictive of future aortic aneurysm (C-index=0.83; 95% CI=0.79,

0.87)), a condition with well-established CT imaging indications and overt structural vascular features readily captured using CT[27,28]. Percival also demonstrated strong performance for predicting myocardial infarction (C-index=0.74; 95% CI=(0.72, 0.77); enrichment=4.62), which may be informed by diffuse radiodensity patterns associated with underlying cardiac or systemic pathology[29,30]. Interestingly, Percival was further capable of predicting future atherosclerosis (**Fig. 5G**; C-index=0.75; 95% CI=(0.74, 0.75); enrichment=3.70) suggesting that it may be identifying subclinical indications of future fat attenuation patterns.

Among prognostic risk signatures for predicting future neoplasms (**Fig. 5C**), 71 demonstrated significant prognostic stratification (mean C-index=0.68; range=0.54, 0.86). Notably, Percival achieved strong performance (C-index=0.76; 95% CI=(0.69, 0.83)) and high disease enrichment (5-fold) for the prediction of future pancreatic cancer, suggesting that Percival captures structural features relevant to early neoplastic changes.

Across a broader set of 118 disease states, Percival demonstrated significant risk stratification within the digestive disease category (**Fig. 5D**; mean C-index=0.64; range=(0.48, 0.78). Within genitourinary diseases (**Fig. 5E**), Percival demonstrated prognostic value for chronic kidney disease (**Fig. 5G**; C-index=0.75; 95% CI=(0.74, 0.76); enrichment=6.02), suggesting that Percival captures subclinical parenchymal or tissue changes. In the musculoskeletal domain (**Fig. 5F**), prediction of future osteoporosis (C-index=0.76; 95% CI=(0.74, 0.78); enrichment=6.02) also reached significance (**Fig. 5G**), reflecting the Percival's sensitivity to subtle morphological indicators of bone health, possibly driven by changes in attenuation.

DISCUSSION

This work introduces Percival, a generalizable CT foundation model trained on over 400,000 CT volumes from over 50,000 participants across multiple contrast protocols, anatomical regions, and associated radiology reports and validated on over 20,000 participants. The cohort was sourced from the PMBB, a large multi-hospital biobank encompassing urban and rural populations, providing geographic and demographic diversity. A key contribution is the scale and clinical diversity of the training data,

enabling the model to generalize across a wide range of imaging and phenotypic clinical contexts. We evaluated Percival's ability to extract meaningful imaging features and clinical associations in four domains: (**1**) zero-shot image-text retrieval to assess alignment of CT volumes with radiology reports and the impact of multi-contrast, multi-view pretraining; (**2**) association analyses to evaluate how Percival's latent feature space aligns with biological markers and clinical measurements; (**3**) development of 307 disease phenotype classification models, including external validation; and (**4**) construction of 678 prognostic model signatures using Cox proportional hazards regression to evaluate the Percival's ability to forecast disease onset and support risk stratification.

Previous efforts to develop medical image foundation models have been primarily focused on two-dimensional data[16,31,32]. While these models benefit from large public datasets, reduced storage, and well-established architectures, they are limited in capturing volumetric, spatial, and anatomical context. In contrast, three-dimensional tomographic imaging data offers rich structural information but presents greater computational and data-related challenges, including larger input sizes, variable scanning protocols, and limited public availability of paired annotations. Recent efforts have made important strides in developing VLMs for CT[17,18]. Merlin, a CLIP-based model trained on 15,000 abdominal CT volumes with paired radiology reports, used a pretrained ResNet image encoder and BERT-style text encoder[17]. Their study demonstrated improved performance across a variety of downstream tasks as well as the ability to perform zero-shot classification. Notably, they further evaluated the prognostic utility of their image encoder by predicting five-year disease outcomes. In contrast, we identify clinically meaningful risk strata to characterize individuals with elevated or reduced risk and enabling more nuanced stratification. CT-CLIP was trained on 50,000 thoracic CT volumes from 21,000 patients[18]. Their study focused on classifying thoracic abnormalities using radiology reports, rather than leveraging structured disease labels such as ICD codes for disease state classification. While these studies represent important progress, they remain limited in scale and scope. Foundation models require large, heterogeneous datasets to enable broad downstream applicability with minimal supervision. In contrast, existing CT-based models have relied

on smaller cohorts focused on specific anatomical regions or contrast protocols, limiting generalizability.

A key contribution of this work is Percival's use of multi-contrast, multi-view contrastive pretraining, enabling the model to learn generalizable, latent feature representations across varying imaging protocols and anatomical FOVs. Percival consistently outperformed architectures trained on fixed FOVs and contrast types, suggesting that training with anatomical redundancy across scan types better captures the continuity of structures such as the aorta, lungs, and spine. Moreover, the inclusion of diverse spatial and contrast contexts may help the model develop more invariant and generalizable anatomical representations. Specifically, the overlap of anatomical content across thoracic, abdominal, and pelvic volumes provides a natural opportunity for the model to develop rich internal representations. Moreover, UMAP dimensionality reduction revealed that Percival organizes volumetric data according to key anatomical and acquisition-specific attributes. This structured representation suggests that Percival effectively distinguishes between different imaging studies, capturing both anatomical and modality-dependent variation. Such clustering likely contributes to cross-modal retrieval and classification performance by enabling precise alignment between image-text latent features.

We further conducted a LabWAS between PCs derived from Percival's latent feature embeddings and temporal laboratory measurements, uncovering a rich alignment between Percival's latent feature space and biological markers across renal, hepatic, metabolic and inflammatory systems. Percival showed consistent associations with renal function indicators such as BUN, creatinine, and eGFR, which reflect glomerular and tubular health. Significant associations were also observed with liver and biliary markers, including ALP, total bilirubin, and urobilinogen, consistent with the Percival's ability to capture features relevant to cholestasis and hepatic function. Moreover, PCs correlated with markers of metabolic health and lipid regulation, including glucose, HbA1c, cholesterol subtypes, triglycerides, and iron, as well as inflammatory and coagulation markers like CRP, ESR, PT, and PTT, which are linked to cardiovascular risk and systemic inflammation. These findings highlight that Percival's

latent space encodes clinically meaningful, biologically grounded information across diverse physiological systems.

To assess the clinical utility of Percival's latent feature space and to inform downstream diagnostic and prognostic modeling, we performed a PheWAS, revealing significant associations across multiple organ systems, improved performance over baseline models, and strong generalization on an external dataset. Percival performed well on disease states with clear CT indications, such as aortic aneurysms, coronary atherosclerosis and cardiomegaly, which have well-established imaging hallmarks, including aortic dilation, vascular calcification, and an enlarged cardiac silhouette. Notably, Percival also performed well on conditions that are not typical clinical indications for CT, including osteoporosis, HFrEF, and atrial fibrillation. These diseases are more commonly assessed using modalities such as echocardiography, ECG, or dual-energy X-ray absorptiometry. Percival's ability to predict these conditions suggests that they may co-occur with or be reflected in other recognizable imaging phenotypes. For example, osteoporosis is associated with subtle changes in bone radiodensity, HFrEF is often associated with cardiomegaly and in some cases structural cardiac remodeling or pulmonary congestion, whereas atrial fibrillation is associated with an enlarged atrium and atrial fibrosis[33,34]. These results indicate that Percival may be leveraging overlapping or correlated imaging signatures to infer disease risk, even when the primary condition is not explicitly visible on imaging. This observation suggests the emergence of at least three distinct categories of disease phenotypes in the context of CT-based foundation modeling. The first includes conditions with clear, well-established imaging indications that are routinely diagnosed on CT, such as aortic aneurysm or pneumothorax. The second includes diseases not traditionally considered CT-indicated but that share imaging features with more visible conditions. Examples include osteoporosis and HFrEF, which demonstrated strong model performance despite not being primary CT targets. These conditions may represent colliding phenotypes, where the disease itself is not directly visible but shares imaging features with other conditions that are. The third category includes diseases that lack both direct visibility and overlapping imaging traits. These conditions exhibited reduced performance, suggesting that they are not easily inferred from CT-based representations alone. Together, these

categories reflect the breadth and limitations of the model's latent space and may help guide future applications and areas for model refinement.

Our study has notable strengths and some limitations. Percival represents a substantial advance in scale, with a training population more than twice the size of previous efforts and approximately eight times as many three-dimensional CT volumes. In total, the model was trained on over 400,000 CT volumes paired with radiology reports from more than 50,000 patients, spanning multiple FOVs, contrast types, and imaging contexts. Moreover, Percival was validated on over 100,000 CT volumes from over 20,000 participants. This breadth and diversity form the necessary foundation for robust, transferable medical AI. Additionally, we present the first comparison of three-dimensional CT foundation models. Notably, some radiology reports are associated with multiple CT volumes due to the standard practice of acquiring multiple imaging volumes within a single study. In such cases, radiologists summarize findings across all CT volumes in a single report, resulting in one report being mapped to multiple scans. This could result in some image-text pairs having incomplete mapping of features. Furthermore, the InfoNCE loss function optimizes paired image-text feature similarity while encouraging separation between unpaired samples. However, because most radiology reports describe normal, unremarkable, or negative findings, some negative image-text pairs may still exhibit inherent similarity. This suggests that the model may learn to distinguish between fine-grained variations in normal imaging patterns rather than solely focusing on pathological differences.

Percival offers a foundation for diverse future applications. Near-term directions include adapting the model for disease classification through task-specific fine-tuning and incorporating longitudinal data to model disease progression. Methodological improvements—such as spatially aware loss functions to better preserve anatomical structure—could further enhance performance. Radiology report pre-drafting remains a particularly impactful use case given rising imaging volumes and limited radiologist capacity. We release Percival publicly to support continued progress in vision–language modeling for medical imaging.

In summary, our study highlights the potential of vision-language foundation models for three-dimensional CT imaging. To our knowledge, Percival is the most

comprehensively trained foundation model for CT imaging to date. By releasing Percival publicly, we aim to provide a valuable resource for advancing future applications in precision medicine.

METHODS

Study population

The Penn Medicine BioBank (PMBB) comprises more than 300,000 consenting patients who receive care in the Penn Medicine healthcare system[35]. Patients were enrolled during routine care appointments at a Penn Medicine location during which they provided written consent to either in-person or digitally to participate in the PMBB. This consent enabled the linkage of their electronic health records, including past disease diagnosis, laboratory measurements, and imaging data. PMBB imaging data were first de-identified by a third-party "honest broker," who removed patient-identifying information from both the DICOM metadata and radiology reports (https://www.med.upenn.edu/radar/). For model training, we used over 400,000 CT volumes paired with radiology reports from 51,000 PMBB participants. For validation, we included 55,000 paired CT-report volumes from 7,000 participants and 82,000 unpaired CT volumes from 17,746 participants, resulting in a combined validation set of over 135,000 CT scans from more than 21,000 PMBB participants. The PMBB was approved by the Institutional Review Board at the University of Pennsylvania under IRB protocol 813913. The PMBB data is available under a data use agreement with the University of Pennsylvania, with access governed by HIPAA-compliant restrictions on re-identification, external sharing, and commercial use. This research was approved by the Institutional Review Board at the University of Pennsylvania under IRB protocol 857974.

Image preprocessing

All CT volumes were acquired in a clinical setting using Siemens, GE HealthCare, Philips, Canon, or Hitachi scanners, stored as DICOM data directories, and subsequently converted to NIfTI format using the dcm2niix conversion tool[36]. All CT

volumes were first resampled to a uniform voxel spacing of 1.5 mm (sagittal) × 1.5 mm (coronal) × 3 mm (axial). Image radiodensities were clipped to a Hounsfield unit (HU) range of -1000 to 1000, followed by min-max normalization to a [0, 1] range. Finally, images were spatially padded and center-cropped to standardized dimensions of 256 × 256 pixels in-plane and 128 slices out-of-plane.

VLM Foundation model architecture

Percival is composed of distinct vision and language encoders designed to establish latent feature space alignment between image and text modalities. The vision encoder leverages a pretrained implementation of the Data-Efficient Image Transformer (DEIT) adapted for three-dimensional tomographic imaging data[37]. To effectively process volumetric CT data, the model utilizes three-dimensional patch embeddings, where each patch represents a 64×64×64 voxel segment. From each CT volume, 32 regional patch embeddings are extracted, along with a single global patch embedding, ensuring comprehensive spatial representation.

For the language encoder, Percival employs the Clinical Longformer architecture, which is well-suited for processing lengthy and structured radiology reports and has been integrated into previous CLIP-based image models[17,38,39]. The model was trained using a vectorized latent feature space of size 512, aligning imaging and text embeddings to optimize multimodal understanding. This multimodal learning approach enables Percival to capture complex relationships between CT volumes and radiological descriptions, facilitating its application to downstream tasks such as image-text retrieval and disease classification and prognostication.

Training protocol

Percival is trained using 402,958 paired three-dimensional CT volumes and corresponding radiology reports from 51,240 PMBB participants and is validated on a holdout set of 55,398 CT volumes and paired radiology reports from 7,000 PMBB participants as well as an additional set of 82,514 CT volumes without paired radiology reports from 17,746 PMBB participants. The combined validation set includes 137,912 CT volumes from 21,875 PMBB participants. The vision and language encoders are

jointly trained using a symmetric contrastive InfoNCE loss function, which optimizes the alignment between imaging and text representations[13,40]. This objective function maximizes cosine similarity between paired representations while minimizing similarity between unpaired representations, effectively distinguishing relevant cross-modal relationships. More specifically, given a batch of b image-text input pairs $([(x_{img}^1, x_{txt}^1), (x_{img}^2, x_{txt}^2), \ldots, (x_{img}^b, x_{txt}^b)])$, the vision and language encoders (G_{img}, G_{txt}) compute corresponding latent feature representations $([(z_{img}^1, z_{txt}^1), (z_{img}^2, z_{txt}^2), \ldots, (z_{img}^b, z_{txt}^b)])$. We then construct a similarity matrix $S \in \mathbb{R}^{b \times b}$ where each element $S_{i,j}$ represents the cosine similarity between z_{image}^i and z_{text}^j. The diagonal of S contains the true paired image-text similarities, while the off-diagonal elements represent negative pairs. InfoNCE loss is subsequently computed using a temperature-scaled softmax:

$$\mathcal{L} = -\frac{1}{b}\sum_{i=1}^{b} \log\left(\frac{\exp\left(\frac{S_{i,i}}{\tau}\right)}{\sum_{j=1}^{b} \exp\left(\frac{S_{i,j}}{\tau}\right)}\right),$$

where τ is a temperature hyperparameter that controls the sharpness of the similarity distribution. This loss encourages image-text pairs to have high similarity while simultaneously pushing negative pairs apart. Training was performed using the AdamW optimizer[41,42] with a batch size of 48, $\tau = 0.1$, and an initial learning rate of 5×10^{-5}. All models were trained in parallel with Pytorch using two Nvidia A100 GPUs.

Zero-shot recall performance evaluation

To evaluate Percival's zero-shot recall performance, we assessed its ability to correctly retrieve matching image-text pairs based on cosine similarity within the latent embedding space. In this setup, a single ground-truth image-text pair was passed through the vision and language encoders to generate corresponding feature embeddings. Next, the model was presented with multiple radiology reports, including the correct match alongside k randomly selected non-matching candidates. Top-5 recall accuracy was quantified by measuring whether the true image-text pair ranked among the five highest cosine similarity scores in a k-way comparison (k = 8, 16, 32, 64, 128, 256). In addition, mean recall rank was computed as the average position of the true

image-text pair when all candidate pairs were ranked by cosine similarity. Performance was evaluated within specific FOVs and contrast methods.

UMAP dimensionality reduction

To interrogate the unsupervised clustering structure of Percival's latent feature space, we applied Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction, followed by color-coding based on imaging FOV, contrast method, age and sex to assess unsupervised clustering patterns. Specifically, we extract the image latent features by passing each CT volume in the validation set through the vision encoder, generating a corresponding latent feature vector. These 512-dimensional vectors were then projected into a two-dimensional space using UMAP, allowing us to visually assess how the model organizes imaging data based on latent feature similarities.

Association analyses of Percival's latent space

To assess the biological and clinical relevance of Percival's latent space we performed principal component analysis (PCA) on participant-level latent feature embeddings and retained the top 10 components (PCs). We first performed a univariate regression of each principal component and age (at the time of imaging). We then conducted a regression analysis of clinical variables as well as a laboratory-wide association study (LabWAS) using a half-half discovery replication design to evaluate whether these components captured biologically meaningful variation. Specifically, PCs were first derived in the discovery cohort (n=10,937) and then applied to the replication cohort (n=10,937). For each of the 7 clinical traits and 62 laboratory traits in the PMBB validation cohort (n=21,875), we fit separate linear regression models for each of the 10 PCs, adjusting for age, sex, and the time interval between imaging and laboratory/clinical measurement. Only clinical and laboratory values obtained within 60 days of imaging were included in the analysis. Significance of associations was determined following false discovery rate (FDR) correction at the 5% level. Next, to assess clinical relevance, we performed a phenome-wide association study (PheWAS) using the latent feature space PCs across 641 disease states. Positive diagnoses were derived from ICD-9CM codes that occurred within 60 days of image acquisition,

whereas participants with earlier diagnosis dates were excluded from that analysis. Logistic regression models were used to classify disease states using the 10 principal components, with age and sex included as covariates. Phenome-wide significance was determined following FDR correction at the 5% significance threshold.

Percival classification models

To perform disease state classification, we developed logistic regression models for 307 disease states using the first ten PCs as predictors, with adjustment for age and sex in the complete PMBB validation cohort (n=21,875). Consistent with the PheWAS design, ICD-9CM codes occurring within 60 days of imaging were used to define positive cases. Participants with a prior diagnosis of the disease state under investigation were excluded. Disease states with fewer than 50 positive cases were omitted from the analysis. To evaluate the added predictive value of latent feature space integration, model performance was compared to a baseline model including only age and sex using five-fold cross-validated area under the receiver operating characteristic curve (AUROC). Principal components were first derived in the training fold and then applied to the validation fold to ensure no data leakage. Statistical significance was assessed using the DeLong test, with Bonferroni correction applied for multiple comparisons ($P<0.05/307$).

External validation

The CT-RATE validation dataset of thoracic CT volumes (n=1,304) was used for external validation of overlapping disease classification models[18]. More specifically, four disease-state classifications (cardiomegaly, pneumothorax, arterial wall calcification, and pleural effusion) were present in CT-RATE which overlapped our disease phenotype classification models. To evaluate generalizability, we trained latent feature space classifiers using only the principal components as input features, excluding age and sex as covariates. Performance was assessed using the AUROC.

Percival prognostic risk modeling

In a similar manner, Cox proportional hazards models were constructed using the first ten latent feature space PCs, with adjustment for age and sex, using ICD-9CM codes to define time-to-event outcomes. Model performance was evaluated using five-fold cross-validated concordance index among the PMBB validation cohort (n=21,875) and compared to a baseline model including only age and sex. PCs were first derived within each training fold and later applied to the validation fold to ensure no data leakage. Next, linear predictor values from the fitted Cox models were used to define risk strata by dividing participants into tertiles corresponding to low-, intermediate-, and high-risk groups. Significant risk stratification performance was assessed using the log-rank test Bonferroni adjusted for multiple comparisons ($P<0.05/678$). Additionally, "disease enrichment" was quantified as the ratio of the number of disease outcomes in the high-risk strata divided by the number of disease outcomes in the low-risk strata.

DATA AVAILABILITY

Access to Penn Medicine BioBank data is provided to investigators at the University of Pennsylvania. Percival's pretrained weights, classification and prognostic models have been made available: https://github.com/cams2b/percival

ACKNOWLEDGEMENTS

We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient- participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878.

FUNDING

W.R.W. is supported by NIH grants P41-EB029460, R01 HL169378, R01 HL137984, UL1 TR001878. J.A.C. is supported by NIH grants R01-HL 121510, R33-HL-146390,

R01HL153646, R01-AG058969, 1R01-HL104106, P01-HL094307, R03-HL146874, K24-AG070459 and R56-HL136730. J.G is supported by R01EB031722, R01HL133889. L.S. is supported by P30 AG073105, U01 AG088658 and U01 AG066833. MDR is supported by UL1TR001878

DISCLOSURE

Dr. Chirinos is supported by NIH grants U01-TR003734, U01-TR003734-01S1, UO1-HL160277, R33-HL-146390, R01-HL153646, K24-AG070459, R01-AG058969, R01-HL157108, R01-HL155599, R01-HL104106 and R01HL155764. He has recently consulted for Bayer, Fukuda-Denshi, Bristol-Myers Squibb, Biohaven Pharmaceuticals, Johnson & Johnson, Edwards Life Sciences, Merck, and NGM Biopharmaceuticals. He received University of Pennsylvania research grants from National Institutes of Health, Fukuda-Denshi, Bristol-Myers Squibb, Microsoft and Abbott. He is named as inventor in a University of Pennsylvania patent for the use of inorganic nitrates/nitrites for the treatment of Heart Failure and Preserved Ejection Fraction and for the use of biomarkers in heart failure with preserved ejection fraction. He has received payments for editorial roles from the American Heart Association, the American College of Cardiology, Elsevier and Wiley, and payments for academic roles from the University of Texas, Boston University, and Virginia Commonwealth University. He has received research device loans from Atcor Medical, Fukuda-Denshi, Unex, Uscom, NDD Medical Technologies, Microsoft and MicroVision Medical. The remaining authors have nothing to disclose.

References

1. Smith-Bindman, R. *et al.* Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA* **322**, 843-856 (2019).
2. Koning, H.J.d. *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New England Journal of Medicine* **382**, 503-513 (2020).
3. null, n. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine* **365**, 395-409.
4. Mettler, F.A. *et al.* Patient Exposure from Radiologic and Nuclear Medicine Procedures in the United States: Procedure Volume and Effective Dose for the Period 2006–2016. *Radiology* **295**, 418-427 (2020).
5. Afshari Mirak, S., Tirumani, S.H., Ramaiya, N. & Mohamed, I. The Growing Nationwide Radiologist Shortage: Current Opportunities and Ongoing Challenges for International Medical Graduate Radiologists. *Radiology* **314**, e232625 (2025).
6. Ganeshan, D. *et al.* Burnout in Academic Radiologists in the United States. *Academic Radiology* **27**, 1274-1281 (2020).
7. Siewert, B. *et al.* Seven Challenges in Radiology Practice: From Declining Reimbursement to Inadequate Labor Force: Summary of the 2023 ACR Intersociety Meeting. *Journal of the American College of Radiology* **22**, 129-138 (2025).
8. Tushar, F.I. *et al.* Classification of Multiple Diseases on Body CT Scans Using Weakly Supervised Deep Learning. *Radiology: Artificial Intelligence* **4**, e210026 (2022).
9. Pickhardt, P.J. *et al.* Automated CT biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: a retrospective cohort study. *Lancet Digit Health* **2**, e192-e200 (2020).
10. Pickhardt, P.J. Value-added Opportunistic CT Screening: State of the Art. *Radiology* **303**, 241-254 (2022).
11. Wasserthal, J. *et al.* TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence* **5**, e230024 (2023).
12. Prevedello, L.M. *et al.* Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions. *Radiology: Artificial Intelligence* **1**, e180031 (2019).
13. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. in *International Conference on Machine Learning* (2021).
14. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. *MedCLIP: Contrastive Learning from Unpaired Medical Images and Text*, 3876-3887 (2022).
15. Zhang, S. *et al.* BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. (2023).
16. You, K. *et al.* CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training. *ArXiv* **abs/2310.13292**(2023).
17. Blankemeier, L. *et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography*, (2024).
18. Hamamci, I.E. *et al.* Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. (2024).

19. Consortium, W.G.f.t.C.P. Estimated Glomerular Filtration Rate, Albuminuria, and Adverse Outcomes: An Individual-Participant Data Meta-Analysis. *JAMA* **330**, 1266-1277 (2023).
20. Seki, M. *et al.* Blood urea nitrogen is independently associated with renal outcomes in Japanese patients with stage 3–5 chronic kidney disease: a prospective observational study. *BMC Nephrology* **20**, 115 (2019).
21. Shlipak, M.G. *et al.* Cystatin C versus Creatinine in Determining Risk Based on Kidney Function. *New England Journal of Medicine* **369**, 932-943 (2013).
22. Vagvala, S.H. & O'Connor, S.D. Imaging of abnormal liver function tests. *Clinical Liver Disease* **11**(2018).
23. Siddique, A. & Kowdley, K.V. Approach to a Patient with Elevated Serum Alkaline Phosphatase. *Clinics in Liver Disease* **16**, 199-229 (2012).
24. Sproston, N.R. & Ashworth, J.J. Role of C-Reactive Protein at Sites of Inflammation and Infection. *Frontiers in Immunology* **Volume 9 - 2018**(2018).
25. Erikssen, G. *et al.* Erythrocyte sedimentation rate: a possible marker of atherosclerosis and a strong predictor of coronary heart disease mortality. *European Heart Journal* **21**, 1614-1620 (2000).
26. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**, 837-845 (1988).
27. Posniak, H.V., Olson, M.C., Demos, T.C., Benjoya, R.A. & Marsan, R.E. CT of thoracic aortic aneurysms. *RadioGraphics* **10**, 839-855 (1990).
28. Litmanovich, D., Bankier, A.A., Cantin, L., Raptopoulos, V. & Boiselle, P.M. CT and MRI in Diseases of the Aorta. *American Journal of Roentgenology* **193**, 928-940 (2009).
29. Gosalia, A., Haramati, L.B., Sheth, M.P. & Spindola-Franco, H. CT Detection of Acute Myocardial Infarction. *American Journal of Roentgenology* **182**, 1563-1566 (2004).
30. Shriki, J.E. *et al.* Incidental Myocardial Infarct on Conventional Nongated CT: A Review of the Spectrum of Findings With Gated CT and Cardiac MRI Correlation. *American Journal of Roentgenology* **198**, 496-504 (2012).
31. Kim, C. *et al.* Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine* **30**, 1154-1165 (2024).
32. Chen, Z. *et al.* A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation. (2024).
33. Vadera, S., Osborne, T., Shah, V. & Stephenson, J.A. Opportunistic screening for osteoporosis by abdominal CT in a British population. *Insights into Imaging* **14**, 57 (2023).
34. Wożakowska-Kapłon, B. Changes in left atrial size in patients with persistent atrial fibrillation: a prospective echocardiographic study with a 5-year follow-up period. *International Journal of Cardiology* **101**, 47-52 (2005).
35. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *Journal of Personalized Medicine* **12**, 1974 (2022).

36. Li, X., Morgan, P.S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods* **264**, 47-56 (2016).
37. Touvron, H. *et al.* Training data-efficient image transformers & distillation through attention. in *International Conference on Machine Learning* (2020).
38. Li, Y., Wehbe, R.M., Ahmad, F.S., Wang, H. & Luo, Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *ArXiv* **abs/2201.11838**(2022).
39. Li, Y., Wehbe, R.M., Ahmad, F.S., Wang, H. & Luo, Y. A Comparative Study of Pretrained Language Models for Long Clinical Text. *Journal of the American Medical Informatics Association : JAMIA* (2022).
40. Oord, A.v.d., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. *ArXiv* **abs/1807.03748**(2018).
41. Kingma, D.P. & Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* **abs/1412.6980**(2014).
42. Loshchilov, I. & Hutter, F. Fixing Weight Decay Regularization in Adam. *ArXiv* **abs/1711.05101**(2017).
43. NIAID VIsual & Medical Arts. (NIAID NIH BIOART, 2024).

Figure legends

Figure 1. Study overviews. (**A**) Overview of Penn Medicine Biobank (PMBB) imaging data pipeline. (**B**) Percival architecture and training protocol. (**C**) Overview of image-text recall class via comparing cosine similarity. (**D**) Latent feature space enrichment analysis, including alignment with biological measurements, disease classification without additional pretraining, and prognostic risk stratification.

Figure 2. Top-5 image-text recall accuracy stratified by field of view (FOV), contrast method, and number of comparisons (64, 128, and 256). (**A-C**) Radar plots comparing recall accuracy of Percival with FOV-specific models (FOV-Thorax and FOV-AbdPel) across contrast protocols (non-contrast, IV contrast, CT angiogram) and anatomical fields of view (thorax, abdomen/pelvis). (**D-F**) Radar plots comparing recall accuracy of Percival and the Merlin model across the same contrast protocols and anatomical regions. (**G**) Mean rank of the correct image-text pair in 128-way recall tasks. Results are shown for the general-purpose Percival model, FOV-specific models (FOV-Thorax and FOV-AbdPel), and the convolutional-based Merlin model, across varying contrast protocols and FOVs.

Figure 3. Percival latent space principal component associations. (**A**) Regression plot of age versus latent-PC 1. (**B**) Heatmap visualization of a half-half discovery-replication design laboratory wide association study of the first five latent-PCs. X-mark indicates significant association following FDR correction at the 5% threshold in both the discovery and replication cohorts. The marks I and O indicate a significant association in the discovery and replication cohorts respectively. (**C**) Circular Manhattan plot depicting phenome-wide associations of Percival shared representation space principal components (latent-PCs). The dotted red line indicates phenome-wide significance ($-\log_{10}(P)$) after adjusting for FDR corrections at the 5% threshold. Upward facing triangles indicate positive relationship with disease whereas downward facing triangles reflect a negative relationship with disease.

Figure 4. Phenome-wide disease classification models using Percival's latent space. (**A**) Five-fold cross validated area under the receiver operating characteristic curve (AUROC) and 95% confidence interval for classifying select disease states across seven disease categories. (**B**) Heatmap visualization of base model (age and sex) and Percival's AUC performance in classifying respiratory disease states. (**C**) Heatmap visualization of base model and Percival's AUC performance in classifying circulatory system disease states. (**D**) Heatmap visualization of base model and Percival's model AUC performance in classifying neoplasm disease states. (**E**) Heatmap visualization of base model and Percival's model AUC performance in classifying digestive disease states. (**F**) Heatmap visualization of base model and Percival's model AUC performance in classifying genitourinary disease states. (**G**) Heatmap visualization of base model and Percival's model AUC performance in classifying musculoskeletal disease states. X indicates significant performance improvement (P<0.05/307). (**H**) Comparison of five-fold internal cross validation and external testing on the CT-RATE dataset across four overlapping disease states. Human anatomical art provided by NIH BIOART[43].

Figure 5. Prognostic utility of Percival's latent feature space via Cox proportional hazards modeling. (**A**) Heatmap visualization of a base model (age and sex) and Percival's model C-index performance for predicting future respiratory disease states. (**B**) Heatmap visualization of base model and Percival's model C-index performance for predicting future circulatory system disease states. (**C**) Heatmap visualization of base model and Percival's model C-index performance for predicting future neoplasm disease states**.** (**D**) Heatmap visualization of base model and Percival C-index performance for predicting future digestive disease states. (**E**) Heatmap visualization of base model and Percival's model C-index performance for predicting future genitourinary disease states. (**F**) Heatmap visualization of base model and Percival C-index performance for predicting future musculoskeletal disease states. X indicates significant risk-stratification performance evidenced by log-rank test Bonferroni corrected for multiple comparisons (P<0.05/607). (**G**) Tertile-based risk stratification using Cox linear predictor across four disease states: (1) coronary atherosclerosis, (2)

chronic kidney disease, (3) osteoporosis, and (4) pleural effusion. Human anatomical art provided by NIH BIOART[43].

Figures

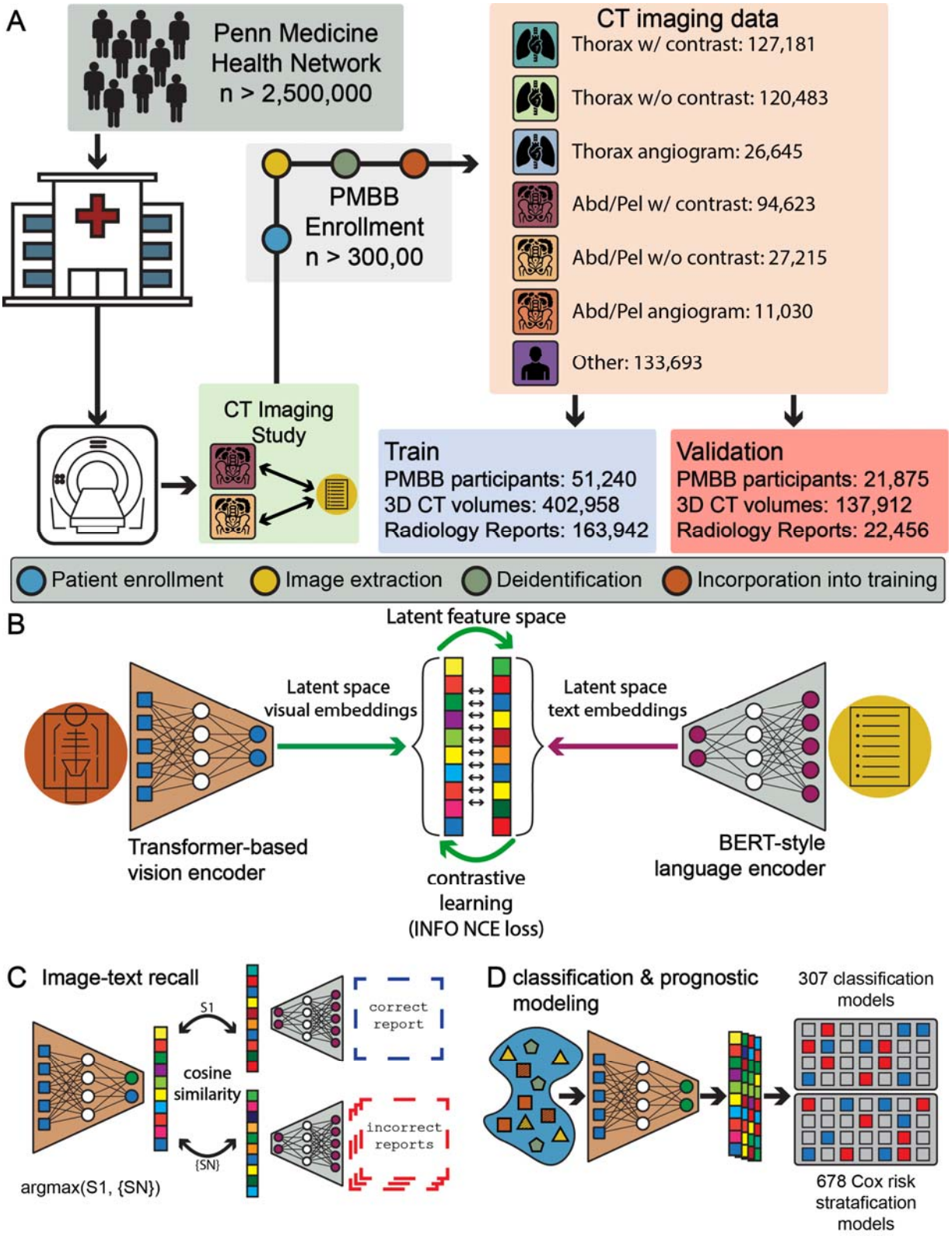


Figure 1.

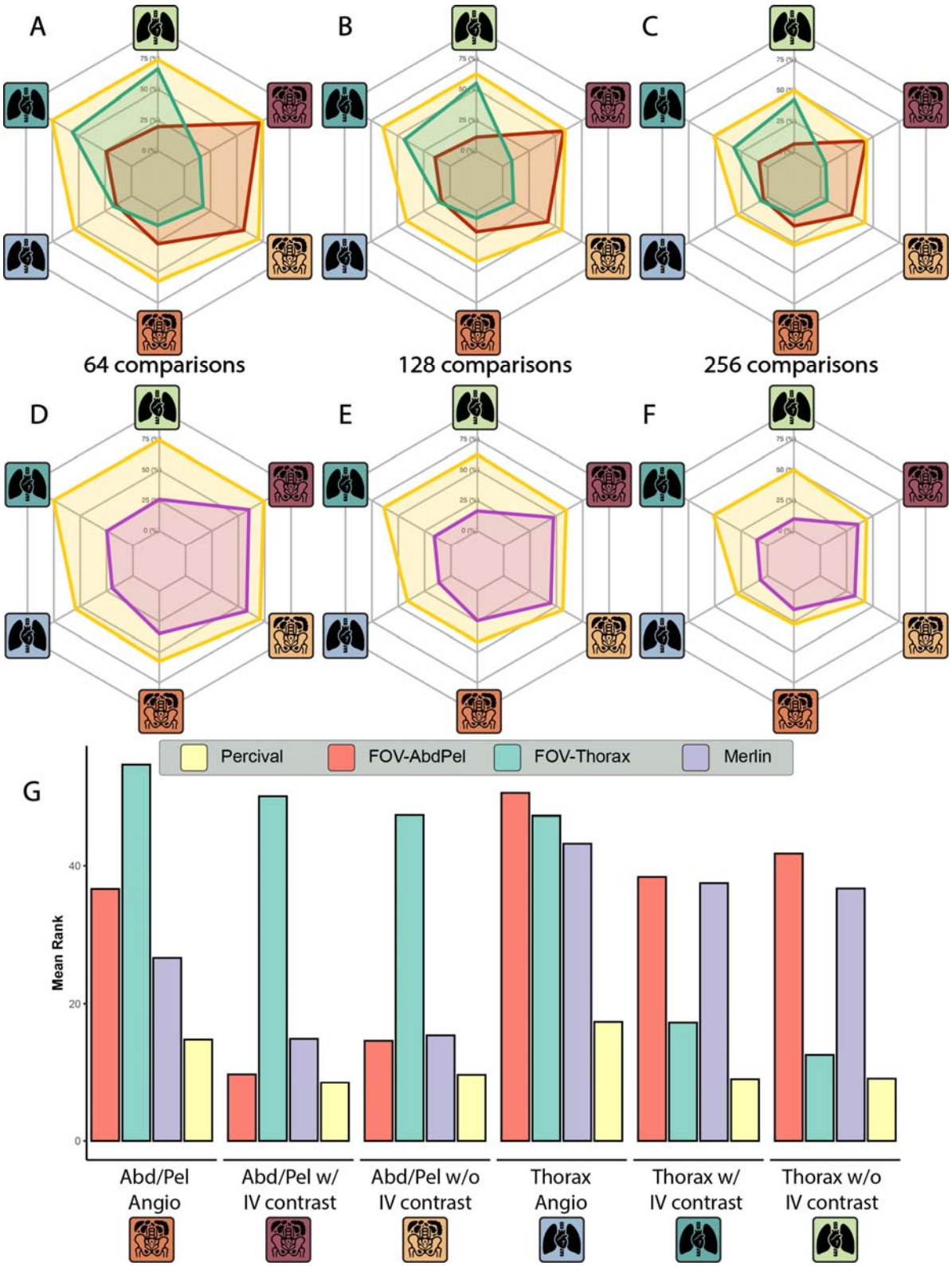

A
B
C
64 comparisons
128 comparisons
256 comparisons
D
E
F
G
Percival
FOV-AbdPel
FOV-Thorax
Merlin
Mean Rank
0
20
40
Abd/Pel Angio
Abd/Pel w/ IV contrast
Abd/Pel w/o IV contrast
Thorax Angio
Thorax w/ IV contrast
Thorax w/o IV contrast

Figure 2.

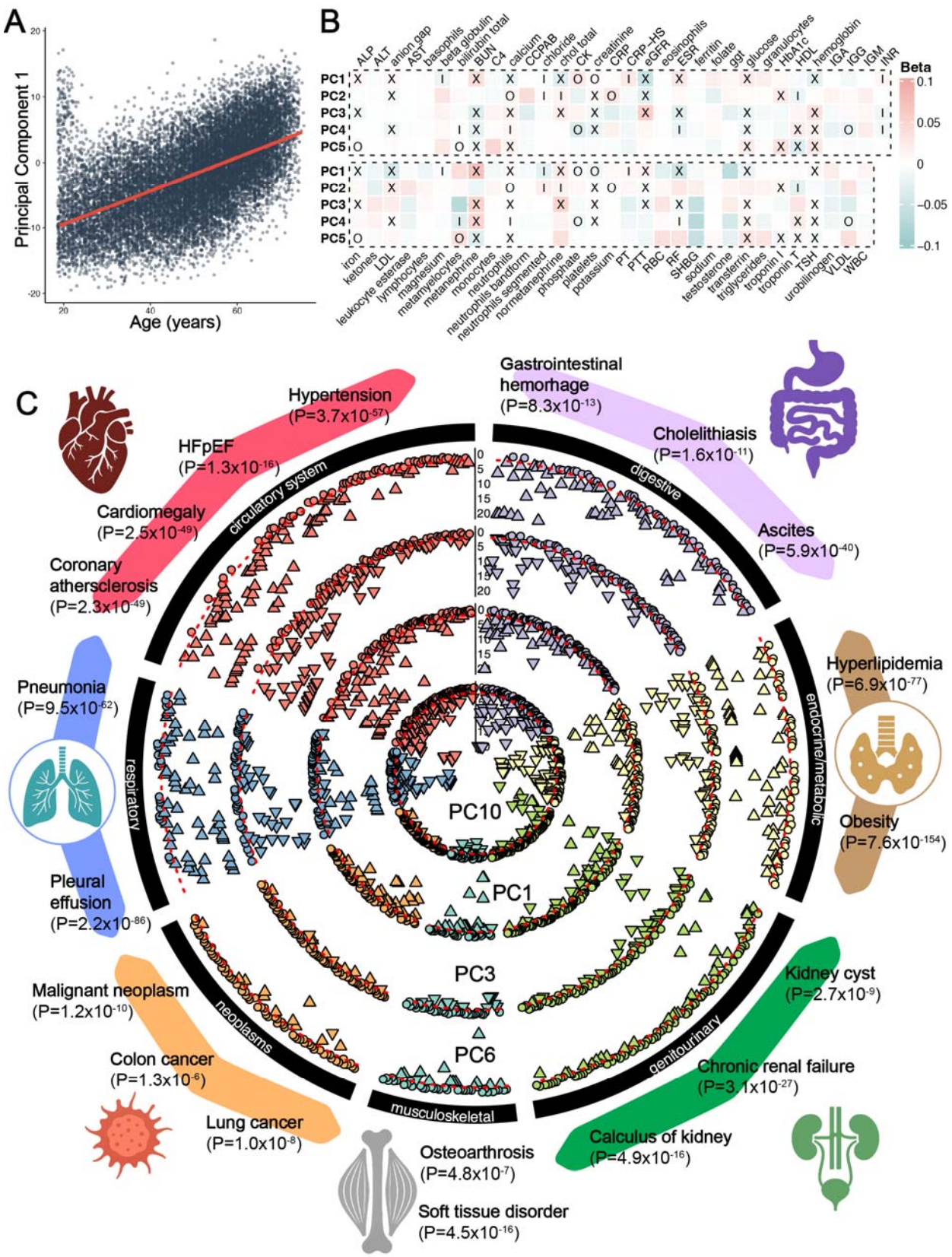
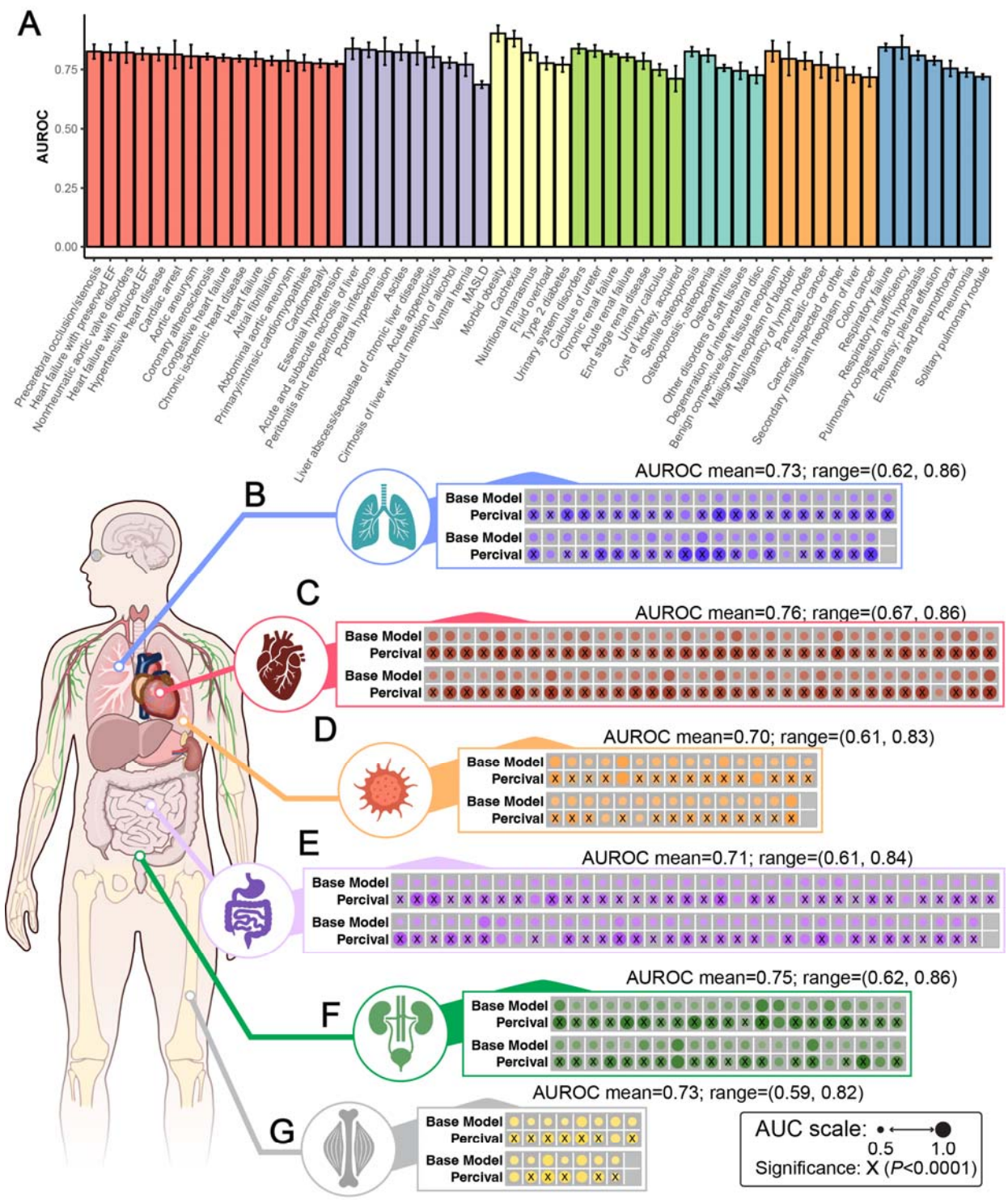

A
Principal Component 1
Age (years)
B
ALP
ALT
anion gap
AST
basophils
beta globulin
bilirubin total
BUN
C4
calcium
CCPAB
chloride
chol total
CK
creatinine
CRP
CRP-HS
eGFR
eosinophils
ESR
ferritin
folate
ggt
glucose
granulocytes
HbA1c
HDL
hemoglobin
IGA
IGG
IGM
INR
PC1
PC2
PC3
PC4
PC5
Beta
iron
ketones
LDL
leukocyte esterase
lymphocytes
magnesium
metamyelocytes
metanephrine
monocytes
neutrophils
neutrophils bandform
neutrophils segmented
normetanephrine
phosphate
platelets
potassium
PT
PTT
RBC
RF
SHBG
sodium
testosterone
transferrin
triglycerides
troponin I
troponin T
TSH
urobilinogen
VLDL
WBC
C
Hypertension
(P=3.7x10-57)
HFpEF
(P=1.3x10-16)
Cardiomegaly
(P=2.5x10-49)
Coronary
atherosclerosis
(P=2.3x10-49)
circulatory system
Gastrointestinal
hemorhage
(P=8.3x10-13)
Cholelithiasis
(P=1.6x10-11)
Ascites
(P=5.9x10-40)
digestive
Hyperlipidemia
(P=6.9x10-77)
Obesity
(P=7.6x10-154)
endocrine/metabolic
Kidney cyst
(P=2.7x10-9)
Chronic renal failure
(P=3.1x10-27)
Calculus of kidney
(P=4.9x10-16)
genitourinary
Osteoarthrosis
(P=4.8x10-7)
Soft tissue disorder
(P=4.5x10-16)
musculoskeletal
Lung cancer
(P=1.0x10-8)
Colon cancer
(P=1.3x10-6)
Malignant neoplasm
(P=1.2x10-10)
neoplasms
Pleural
effusion
(P=2.2x10-86)
Pneumonia
(P=9.5x10-62)
respiratory
PC10
PC1
PC3
PC6

Figure 3.



H

	Arterial wall Calcification	Cardiomegaly	Emphysema	Pleural effusion
CT RATE external	**0.81**	**0.83**	**0.72**	**0.84**
PMBB 5–fold internal	**0.79**	**0.78**	**0.75**	**0.79**

Figure 4.

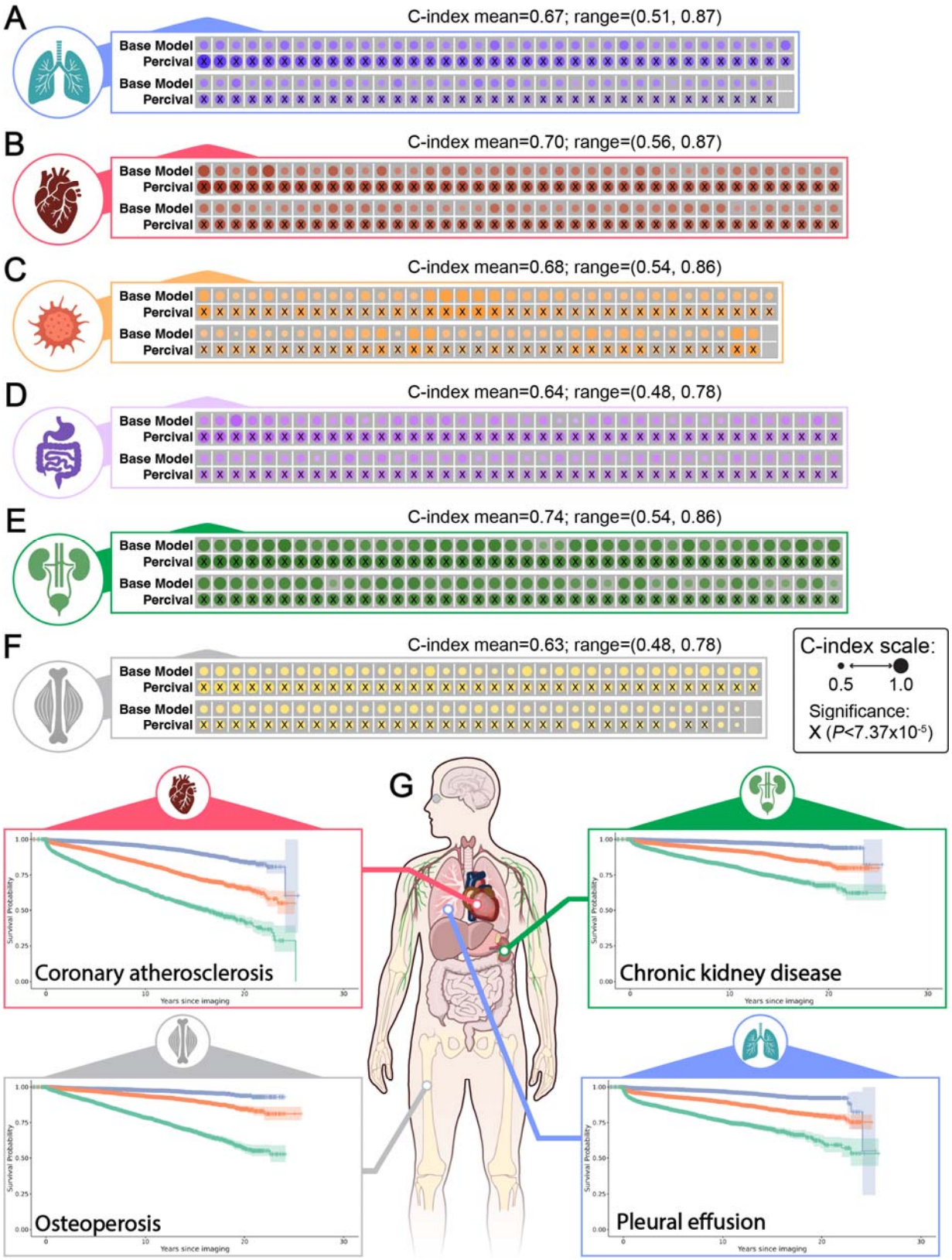

A
C-index mean=0.67; range=(0.51, 0.87)
B
C-index mean=0.70; range=(0.56, 0.87)
C
C-index mean=0.68; range=(0.54, 0.86)
D
C-index mean=0.64; range=(0.48, 0.78)
E
C-index mean=0.74; range=(0.54, 0.86)
F
C-index mean=0.63; range=(0.48, 0.78)
Base Model
Percival
C-index scale:
0.5
1.0
Significance:
X (P<7.37x10-5)
G
Coronary atherosclerosis
Chronic kidney disease
Osteoperosis
Pleural effusion
Survival Probability
Years since imaging

Figure 5.