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

Cameron A. Beeche, BS[1,2,3]; Joonghyun Kim[1]; Hamed Tavolinejad, MD[2,4]; Bingxin Zhao, PhD[5]; Jessie Dong, BS[1];Rakesh Sharma, BS[1]; Jeffrey Duda, PhD[3]; James Gee, PhD[3]; Farouk Dako, MD[3]; Anurag Verma, PhD[6,7]; Colleen Morse MS[6,7]; Bojian Hou, PhD[8]; Li Shen, PhD[8]; Hersh Sagreiya, MD[3]; Christos Davatzikos, PhD[9]; Scott Damrauer, MD[10,11,12]; Rohan Shad, MD[13]; Marylyn D. Ritchie, PhD[10]; Daniel Rader, MD[6,7,10]; Qi Long, PhD[8]; Eric Eaton, PhD[14]; Tianlong Chen, PhD[15]; Charles E. Kahn, Jr., MD, MS[3]; Julio Chirinos, MD, PhD[2,3]; Walter R. Witschey, PhD[3*]; 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]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

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

[5]Department of Statistics and Data Science, 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 Surgery, Division of Cardiovascular Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.
[14]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 19104, USA.
[15]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

Vision-language foundation models (VLMs) for computed tomography (CT) are emerging tools capable of learning generalizable representations from large-scale clinical imaging data. Yet, it remains unclear to what extent these models encode biologically meaningful information relevant to real-world clinical variation. We introduce Percival, a CT-native VLM trained on more than 400,000 CT-report pairs from the Penn Medicine BioBank using a dual-encoder symmetric contrastive framework, with the objective of characterizing the biological associations embedded through contrastive pretraining. Across over 20,000 held-out participants, Percival's latent space shows strong alignment with clinical attributes, body-size measures, and multiple laboratory biomarkers. Phenome-wide analyses further reveal broad correspondence between latent features and disease phenotypes, including conditions not typically evaluated by CT; survival analyses demonstrate that the embeddings capture longitudinal risk patterns. Together, these findings reveal that CT-VLMs uncover a rich latent structure aligned with physiological measurements and disease phenotypes spanning the disease-prevalence spectrum.

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

INTRODUCTION

Computed tomography (CT) is widely used for disease diagnosis and monitoring[1-3], but rising utilization, radiology workforce shortages, and access disparities[4-6] are driving the need for AI-powered tools to improve efficiency and support precision medicine[7-14]. Self-supervised vision language models (VLMs), trained using CT volumes and accompanying radiology reports, have been proposed to address these issues when fine-tuned for downstream tasks such as radiology report drafting[15,16]. However, it remains unclear whether these models encode biologically meaningful variation and whether this extends beyond findings typically associated with CT indications. Existing evaluations have focused largely on computational tasks of classification and report generation, offering limited insight into whether CT-VLMs learn representations that align with patient physiology, laboratory biomarkers, multisystem disease burden, or longitudinal risk. Moreover, CT-VLMs have not yet been examined within large, real-world clinical biobanks to understand the extent to which their learned embeddings capture information beyond common CT indications or whether they reflect patterns spanning the full disease-prevalence spectrum, including heterogeneous, systemic, and low-prevalence conditions. Clarifying these properties is important for determining how a CT-VLM represents underlying biology and disease processes, and for assessing its potential to serve as a broadly informative foundation model.

Leveraging the Penn Medicine BioBank (PMBB), we developed Percival, a vision–language foundation model for three-dimensional CT imaging. Percival was trained on more than 400,000 CT volumes paired with corresponding radiology reports collected from multiple hospitals spanning the Penn Medicine Healthcare system, providing the scale and heterogeneity needed to interrogate how CT-based vision-language models encode biologically meaningful variation beyond standard imaging indications. This multi-site dataset captures substantial diversity in imaging protocols, scanner types, contrast usage, and patient populations, encompassing thoracic, abdominal, pelvic, head and neck, brain, and extremity CT examinations. All radiology reports were authored by licensed physicians and finalized by board-certified radiologists and were deidentified of patient health information, ensuring clinically verified text aligned with real-world diagnostic practice. To our knowledge, this study constitutes the first biobank-scale

characterization of a three-dimensional CT-VLM's latent representation space, systematically linking learned embeddings to lab-wide biomarkers, phenome-wide ICD diagnoses, longitudinal prognostic outcomes, and multisystem disease burden using pan-organ, multi-contrast, multi-site clinical CT data.

In this study, we sought to determine the extent to which a CT-based VLM encodes biologically meaningful variation within a large clinical biobank. We first examined whether Percival's latent representations capture fundamental physiological and clinical attributes by quantifying their associations with demographic factors, body-size measures, and laboratory biomarkers, and by testing whether these embeddings support the generation of synthetic laboratory values. Second, we evaluated phenotypic alignment by mapping phenome-wide disease associations, characterizing correspondence with multisystem disease burden, and assessing whether the learned representations capture conditions spanning the full disease-prevalence spectrum—including heterogeneous, systemic, and low-prevalence diseases not typically indicated for CT imaging. Finally, we assessed the clinical utility of these embeddings for longitudinal risk stratification by testing whether Percival's latent features predict future disease trajectories and prognostic outcomes across 7 ICD chapters. An overview of the study design is presented in **Figure 1**. To support continued research, we publicly release Percival's pretrained weights and accompanying code (https://github.com/cams2b/percival).

RESULTS

Here, we present Percival, a contrastive vision–language model for three-dimensional CT imaging trained in a manner consistent with recent CT-VLM efforts such as Merlin and CT-CLIP, pairing volumetric CT studies with their accompanying radiology reports to learn shared visual–textual representations[15,16]. The study cohort included 73,115 PMBB participants (51,240 for training and 21,875 for validation) with comparable demographic distributions (**Table S1**). The training subset contributed 402,958 CT volumes spanning diverse anatomical regions, contrast phases, scanner types, and acquisition protocols across multiple clinical sites (**Tables S2-S3**). To verify Percival's out-of-domain generalization, we performed zero-shot classification on the external CT-RATE dataset, where the model achieved performance comparable to in-domain CT-CLIP (**Fig. S1**;

Table S4). Unsupervised visualization of the learned representations further revealed organized clustering by CT field of view, contrast type (**Fig. S2A)**, and demographic attributes (**Figs. S2B–S2J**), indicating that Percival captures meaningful variation present in real-world imaging data.

Biological associations with Percival's latent space

Percival's latent feature embeddings were reduced using principal component (PC) analysis with the first ten PCs explaining over 50% of the latent feature embedding variance (**Fig. S3**). Regression analyses testing associations between these PCs and key physiological traits—age, height (cm), weight (kg), systolic blood pressure (mmHg), and diastolic blood pressure (mmHg)—identified 46 of 50 associations as significant after Bonferroni correction (**Table S5**). The dominant axis (PC1) showed the strongest relationship with age (R^2=0.38; **Fig. 2A**), while height and weight were most correlated with components 3 (R^2=0.19) and 5 (R^2=0.41; **Fig. 2C**), respectively. Although all ten PCs demonstrated statistically significant associations with systolic and diastolic blood pressure, the effect sizes were weak (R^2<0.01), indicating limited blood pressure–related signal is captured in the latent space. Sex differences were evaluated using ANOVA with effect size measured by Cohen's D (**Fig. 2B**; **Table S6**). The leading components (PC1-PC3) demonstrated large effect sizes (Cohen's D>1), reflecting biological and morphological distinctions between males and females captured in CT imaging, such as differences in thoracic cavity size, muscular density, and body composition[17-19]. In contrast, components 4-10 showed minimal sex effects (Cohen's D<0.2), suggesting that these latent features encode shared physiological variation rather than sex-specific patterns.

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. 2D**; **Table S7**). Following FDR correction at the 5% threshold, our discovery cohort (average n=3,265) identified 147 significant associations with 43 laboratory measurements, with 100 associations across 30 lab measurements being further verified in the replication cohort (average n=3,268). 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[20-22]. Several components 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[23,24]. There were several associations with glucose, HbA1c, total cholesterol, HDL, LDL, triglycerides, and iron. These markers reflect cardiovascular and metabolic health, glycemic control, and lipid regulation[25-29]. 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[30,31].

Synthetic lab values

We next generated synthetic laboratory values from Percival's latent feature space to assess the extent to which CT-VLM's feature representations capture biochemical variation. Linear probing was performed using Lasso regression with an 80:20 train–validation split, allowing estimation of laboratory traits in participants without concurrent blood draws (**Fig. 2E**). This approach provides an in-silico approximation of physiological markers directly from CT-derived embeddings. Although overall prediction accuracy was modest (mean validation R^2=0.12), several laboratory traits showed stronger correspondence with the latent space. Albumin, HDL cholesterol, and hemoglobin each achieved validation R^2>0.20, suggesting that aspects of these biochemical phenotypes are reflected in Percival's imaging-based representations. **Figure 2E** presents representative train-validation performance across 12 laboratory panels, with complete results summarized in **Table S8**.

Disease enrichment patterns and lifespan disease burden

To characterize the breadth of disease-related variation represented in Percival's latent space, we conducted a phenome-wide association study (PheWAS) across 641 ICD-9/10CM codes in the PMBB validation cohort (average n = 20,912). This analysis revealed extensive phenotypic alignment between Percival's embeddings and clinical diagnoses, with 1,861 significant associations across 521 disease states after FDR correction (**Fig. 3A, Table S9**). These associations spanned multiple organ systems—including 116 circulatory, 65 respiratory, 60 neoplastic, and 74 endocrine/metabolic conditions—indicating that Percival's latent representations capture a wide range of disease phenotypes beyond the findings typically visible on CT. To assess the longitudinal clinical impact of these phenome-wide associations, we next quantified post-imaging disease accumulation as a function of latent biological features. Each participant's principal component axes were binned using standardized z-score (−3 to +3). For each bin, we measured the mean number of unique ICD-9/10CM diagnoses recorded after the imaging date, defining a prospective measure of lifespan disease burden (**Fig. 3B**). A clear disease accumulation gradient was observed across multiple components, with individuals at the extremes of PC values demonstrating a markedly greater accumulation of new diagnoses across circulatory, metabolic/endocrine, and respiratory systems. Specifically, component 3 exhibited the strongest disease gradient for circulatory diseases, with participants binned at <-3 accruing on average 10.5 circulatory related diseases post imaging (**Fig. 3C**). Similar patterns were observed amongst component 5 for endocrine and metabolic diseases, with participants on either extreme accruing >5 diseases over their lifespan (**Fig. 3D**). These relationships show that the latent space of a CT-VLM can capture multisystem patterns of disease risk and phenotypic variation. Such biological alignment indicates that CT-VLMs may provide general-purpose representations that extend beyond discrete imaging tasks and relate to broader clinical outcomes.

Disease classification

To assess the disease-related signal inherently encoded during pretraining, we applied linear probing across 378 clinical conditions using a fully frozen VLM encoder. Percival's

latent embeddings were reduced to principal components to enable stable evaluation across both common and low-prevalence diseases. For each condition, we fit logistic regression models using these components as predictors and evaluated discriminative ability using five-fold cross-validated AUC. Performance was compared against a baseline model containing age and sex. Predictive value beyond demographic covariates was quantified using the DeLong test[32]. Among 378 disease classification models, sample sizes varied widely, with 50% including fewer than 108 positive disease states (**Fig. 4A**). When performance was examined as a function of disease prevalence, AUC exhibited a moderate, positive, linear relationship (**Fig. 4B**). In total, 254 classification models exhibited significant AUC performance gains following FDR correction at the 5% threshold (**Table S10**). For certain disease states, including osteoporosis and atherosclerosis, Percival does not outperform demographic features, whereas for conditions such as heart failure, chronic liver disease, and renal failure, it demonstrates substantially stronger performance. Specifically, 48 circulatory disease models demonstrated performance gains with Percival (AUC mean=0.74; range=0.61, 0.84). Notably, Percival's prediction of heart failure with reduced ejection fraction (HFrEF) exhibited an AUC of 0.77 (95% CI=0.73, 0.82). Among vascular diseases, Percival's prediction of aortic aneurysm exhibited a five-fold AUC of 0.79 (95% CI=0.77, 0.82) respectively. Strong performance was also observed for non-standard imaging indications such as atrial fibrillation (AUC=0.76; 95% CI=0.74, 0.77). Respiratory diseases demonstrated similar trends, with 39 conditions showing significant classification performance and a mean AUC of 0.72 (range=0.62, 0.81). The model for respiratory failure achieved the highest respiratory performance (AUC = 0.81; 95% CI=0.77, 0.84).

Percival's performance in predicting neoplastic diseases was comparable to that of other categories (mean AUC=0.71; range=0.61, 0.82); 28 neoplasm models demonstrated significant gains relative to the baseline. Notable performance was observed for prediction of suspected cancer (AUC=0.75; 95% CI=0.72, 0.79). Additional improvements were identified across digestive (significant outcomes=61; mean AUC=0.71; range=0.61, 0.83), genitourinary (significant outcomes=26; mean AUC=0.76; range=0.60, 0.83), endocrine/metabolic (significant outcomes=43; mean AUC=0.75; range=0.64, 0.86), and musculoskeletal (significant outcomes=9; mean AUC=0.72;

range=0.66, 0.78) conditions. Among these chapters, several conditions not typically assessed on CT also showed strong predictive signal from the frozen embeddings, including morbid obesity (AUC=0.86), diabetes mellitus (AUC=0.78), and cachexia (AUC=0.82). These phenotypes reflect systemic metabolic and physiologic states that are not directly measured on CT, indicating that CT-VLM's pretrained representations capture broader biological variation beyond visible radiologic findings. AUC performance for select outcomes is presented in **Fig. 4C**. Sex-specific replication of disease classification models was feasible for 175 disease states, with males and females showing 135 and 119 significant performance gains over demographic baselines, respectively (**Table S11**). Overall, AUC performance was largely concordant (Pearson R=0.74) between sexes (**Fig. S5**).

Examination of outcomes across the disease spectrum

To further characterize how CT-VLM representations capture disease-related variation, we next examined outcomes spanning the disease-prevalence spectrum. This analysis enabled us to evaluate whether Percival's latent features distinguish clinical phenotypes that range from rare to common and to explore how representation structure differs across these categories. We selected one low-prevalence condition (respiratory failure), one moderate-prevalence condition (aortic aneurysm), and one high-prevalence condition (chronic kidney disease) to assess how the pretrained embeddings separate affected from unaffected participants and how these patterns reflect underlying biological dynamics across disease types. For each condition, we performed a component analysis by examining the distribution of latent features and visualized group-level differences through density plots and Cohen's D effect size analyses to reveal how disease prevalence relates to the separability and structure of underlying feature representations within Percival's latent space (**Fig. 5**). Respiratory failure, the lowest-prevalence condition included in our analysis, showed strong discriminative performance with a five-fold cross-validated AUC of 0.81 (**Fig. 5A**). Component analysis of the first ten principal components revealed clear separation between affected and unaffected participants, with the largest effect sizes observed for PC6 (Cohen's D=0.65), PC9 (Cohen's D=0.62), and PC3 (Cohen's D=0.53) (**Fig. 5B–C**). Aortic aneurysm, a moderate-prevalence condition

(n=339), achieved an AUC of 0.79 (Fig. 6D), with the greatest separation on PC1 (Cohen's D=0.75) and PC8 (Cohen's D=0.74) (**Fig. 5E–F**). Chronic kidney disease (AUC=0.78; **Fig. 5G**), a high-prevalence condition not typically diagnosed via CT, showed more diffuse separability, with smaller effect sizes across multiple components including PC1, PC3, PC6, and PC9 (Cohen's D range=0.42, 0.62) (**Fig. 5H-I**).

Prognostic modeling

To evaluate whether CT-VLM's pretrained representations capture information relevant to future health trajectories, we examined Percival's prognostic utility across a wide range of clinical outcomes. Using survival analysis, we assessed whether the latent features encode patterns associated with elevated or reduced risk of subsequent disease development (**Fig. 6**; **Table S12**). Our analysis proceeded in two parts. First, we applied Cox proportional-hazards models to 648 disease states and quantified prognostic signal using five-fold cross-validated concordance index (C-index), providing a systematic assessment of the longitudinal information contained in the imaging-derived embeddings. Second, for each disease state, we derived latent feature-based risk stratification profiles that separated participants into low-, intermediate-, and high-risk groups (**Figs. 6A-D**). Differences in survival curves were evaluated using the log-rank test with significance determined by Bonferroni correction ($P<0.05/648$). Among respiratory diseases, Percival demonstrated significant risk stratification across 70 disease states with an overall mean C-index of 0.67 (range=(0.52, 0.86)). Notably, Percival predicted future respiratory failure with a C-index of 0.77 (95% CI=0.75, 0.79; **Fig. 6A**).

Across circulatory diseases, Percival achieved significant risk stratification for 123 conditions (mean C-index=0.70; range=0.56, 0.83), including chronic pulmonary heart disease (C-index=0.76; 95% CI=0.75, 0.78; **Fig. 6B**). Percival also predicted vascular outcomes such as future aortic aneurysm (C-index=0.83; 95% CI=0.79, 0.88), a condition with well-established CT imaging indications and overt structural vascular features readily captured using CT[33,34]. Interestingly, Percival also predicted heart failure with reduced ejection fraction (C-index=0.78; 95% CI=(0.75, 0.82)) which is typically diagnosed via functional imaging.

Within endocrine and metabolic diseases, Percival achieved significant prognostic stratification for 85 conditions (mean C-index=0.68; range=0.54, 0.85), including strong prediction of future cachexia (C-index=0.78; 95% CI=0.74, 0.82; **Fig. 6C**). Percival also stratified 113 genitourinary diseases (mean C-index=0.70; range=0.49, 0.85), highlighted by acute renal failure (C-index=0.75; 95% CI=0.71, 0.80; **Fig. 6D**), suggesting sensitivity to subclinical parenchymal and tissue-level changes. Percival also stratified risk across neoplastic (72 states, mean C-index=0.69), digestive (109 states, mean C-index=0.65), and musculoskeletal diseases (59 disease-states, mean C-index=0.66), indicating broad prognostic sensitivity to structural and tissue-level variation. Sex-specific replication of disease classification models was feasible for 404 disease states, with males and females C-index performance being largely concordant (Pearson R=0.75) between sexes (**Fig. 6F; Table S13**).

DISCUSSION

This work introduces Percival, a large-scale CT vision–language model trained on more than 400,000 CT volumes from over 50,000 participants collected from multiple hospitals spanning the Penn Medicine Healthcare system. Prior work on CT vision-language models has primarily evaluated performance on task-specific benchmarks such as classification, retrieval, or report generation, often with supervised fine-tuning. In contrast, our study focuses on a systematic characterization of the biological, phenotypic, and prognostic information inherently encoded in a pretrained CT-VLM's latent representation space, without task-specific adaptation. The cohort was sourced from the Penn Medicine BioBank, a multi-hospital system serving urban and rural communities, providing substantial geographic, demographic, and clinical diversity. A key contribution of this work is the use of a heterogeneous, real-world biobank to examine not only how CT-VLMs perform on conventional imaging tasks, but more importantly, how pretrained CT-VLM representations reflect underlying human biology and disease. Our evaluation therefore focused on quantifying the biological and phenotypic structure captured within the model's feature space. We first assessed whether Percival encodes core physiological and demographic attributes, including age, sex, body-size measures, blood pressure measurements, and laboratory biomarkers. We then examined broader clinical relevance

through phenome-wide disease associations, characterization of multisystem disease burden, and evaluation across the full disease-prevalence spectrum—from common to rare conditions—to determine whether the learned representations generalize to low-signal clinical settings. Finally, we assessed the prognostic utility of the latent features by modeling longitudinal risk across 7 ICD chapters.

Early progress in medical vision–language modeling has been driven largely by two-dimensional imaging [35-37], where abundant public datasets and mature architectures have enabled rapid methodological advances. 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[15,16]. Merlin, a CLIP-based model trained on abdominal and pelvic CT volumes with paired radiology reports, used a pretrained ResNet image encoder and BERT-style text encoder[15]. Their study demonstrated improved performance across a variety of downstream tasks, suggesting that contrastive pretraining can yield latent features that capture clinically relevant patterns. CT-CLIP extended these ideas to thoracic CT imaging, focusing on thoracic abnormality classification and report generation, while additionally illustrating how unsupervised clustering within the learned feature space may reflect underlying radiologic patterns. Together, these models provide important momentum toward developing foundation models for CT and demonstrate that VLMs can perform well across a range of classification and interpretation tasks.

What remains less understood, however, is the degree to which CT-based VLMs encode the biological and clinical variation necessary for broad applicability in real-world medical settings. Existing evaluations offer limited insight into whether the learned representations capture deeper physiological attributes, laboratory biomarkers, multisystem disease burden, or longitudinal risk. In this work, we address this gap by systematically validating a CT-VLM at biobank scale and quantifying the biological, phenotypic, and prognostic structure encoded within its latent space. Our contribution lies in characterizing how a CT-VLM functions as a scientific model of human biology and

disease—providing a crucial step toward understanding the clinical readiness and foundational properties of VLMs for medical imaging.

Percival's multi-contrast, multi-view contrastive pretraining enabled the model to learn latent representations that generalize across diverse imaging protocols and anatomical fields of view. On out-of-domain classification tasks (CT-RATE), Percival performed comparably to existing CT-VLM architectures, demonstrating reliable generalization without task- or domain-specific tuning. Importantly, exposure to diverse anatomical coverage across thoracic, abdominal, and pelvic CT volumes enables the model to learn coherent representations of continuous biological structures—such as the aorta, lungs, and spine—that extend across multiple contiguous regions and are only partially observed within any single field of view. Variation in spatial context and contrast phase further exposes the model to physiologic differences across organ systems, supporting the learning of more biologically grounded and transferable feature representations.

Evaluation of Percival's latent features showed that the leading principal components capture key biological attributes—including age, sex, height, and weight—while subsequent components represent physiological variation common across individuals. We further conducted a LabWAS between PCs derived from Percival's feature space 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[20,21,38,39]. Significant associations were also observed with liver and biliary markers, including ALP, total bilirubin, and urobilinogen, consistent with 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. We also evaluated Percival's capacity to generate synthetic laboratory values from its latent embeddings. Although performance was modest, the results indicate that CT-VLM feature representations retain measurable biochemical information. These

synthetic estimates are not intended to replace clinical laboratory testing, but they offer an in-silico approximation of physiological status and may help identify individuals with extreme or unexpected values who warrant additional triage or follow-up. These findings highlight that Percival's latent space encodes clinically meaningful, biologically grounded information across diverse physiological systems.

To examine how a CT-VLM aligns with clinically meaningful disease phenotypes, we evaluated associations between the latent feature space and diagnoses spanning 7 ICD chapters. Diseases with clear anatomical correlates on CT—such as aortic aneurysm and cardiomegaly—showed strong alignment, consistent with their distinct structural manifestations (e.g., aortic dilation or enlargement of the cardiac silhouette). Notably, the latent representations also captured variation across the full disease-prevalence spectrum, with only modest increases in performance as prevalence increased. This suggests that CT-VLMs encode features that generalize beyond high-signal imaging findings and retain clinically relevant information even for conditions with limited case numbers. Given that nearly half of all evaluated diseases had fewer than 100 positive cases, these results highlight the importance of developing VLMs that remain informative in data-sparse clinical settings. Component analysis across representative diseases further illustrated how prevalence relates to the structure of latent feature differences. The lowest-prevalence condition, respiratory failure, showed its strongest separability on higher-order, lower-variance components, indicating that rare diseases may be encoded through more subtle latent dimensions. Aortic aneurysm, a moderate-prevalence condition with overt CT findings, demonstrated its largest effects on both the dominant axis (PC1) and a mid-range component (PC8), reflecting the combination of global and localized structural changes associated with aneurysmal dilation. In contrast, chronic kidney disease, a highly prevalent and non-CT-indicated condition, showed more diffuse and modest differences across several components. These patterns highlight the importance of models that can capture both prominent global axes of variation and the subtler signals that characterize rare or systemic diseases.

Percival also showed meaningful associations for conditions that are not typically evaluated using CT, including chronic kidney disease, heart failure with reduced ejection fraction, and atrial fibrillation, despite these diagnoses relying primarily on laboratory

assessment (e.g., eGFR for CKD), alternative imaging modalities (e.g., echocardiography for HFrEF), or electrocardiographic evaluation (e.g., ECG for atrial fibrillation). Percival's ability to predict these conditions suggests that they may co-occur with or be reflected in other recognizable imaging phenotypes. For example, 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[40,41]. These findings suggest that CT-based VLMs can leverage overlapping or correlated imaging signatures to capture disease-related information even when the primary condition is not directly visible on CT. More broadly, they point toward the emergence of distinct categories of disease phenotypes in the context of CT-VLMs, reflecting differences in how biological processes manifest in imaging data. 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 CKD 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, comparable to patient demographics, suggesting that they are not easily inferred from CT-based representations alone. Together, these categories illustrate both the breadth and the inherent limits of the biological signal captured in CT-VLM embeddings, clarifying how such models encode different dimensions of disease-related information.

Our prognostic analyses further demonstrated that CT-VLM's latent representation space captures variation that is predictive of future multisystem morbidity. Participants with extreme values along specific principal component axes accumulated more post-imaging diagnoses across circulatory, metabolic, endocrine, and respiratory systems, indicating that certain latent dimensions reflect underlying biological vulnerability that manifests across organ systems over time. Consistent with these gradients, survival modeling revealed that Percival's representations stratified risk across a wide range of

disease states, including both conditions with clear CT correlates and others not typically diagnosed through CT, suggesting sensitivity to structural or tissue-level patterns that precede clinical presentation. Importantly, these prognostic signals emerged without task-specific supervision, indicating that Percival encodes features tied not only to present disease states but also to downstream risk trajectories. This property provides a foundation for using CT-VLM representations to study multisystem health trajectories and informs subsequent efforts to fine-tune models for time-to-event outcomes or individualized risk stratification.

This study has notable strengths and some limitations. Percival was developed within a large clinical biobank of more than 70,000 participants with linked imaging and longitudinal health records. Over 50,000 participants contributed 400,000 CT volumes paired with radiology reports for training, spanning diverse fields of view, contrast phases, and acquisition settings. Validation on an additional 20,000 participants with more than 100,000 CT volumes enabled a robust assessment of generalizability across real-world imaging contexts. Notably, radiology reports are commonly associated with multiple CT volumes. 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, Percival's VLM was trained using the InfoNCE loss function, which 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.

Future extensions of CT-based vision–language models will benefit from deeper integration of biological information alongside imaging. Explicit incorporation of laboratory measurements, longitudinal clinical trajectories, and, where available, genomic or biomarker data could help illuminate how diverse biological processes are reflected in CT-derived representations. Linking multimodal biological signals to latent imaging features may also clarify which aspects of systemic disease are captured by VLMs and which remain outside the scope of tomographic imaging. Beyond phenotype prediction,

these models hold promise for applications such as automated report pre-drafting and clinical decision support, particularly in settings with rising imaging volumes and limited radiologist capacity. We release Percival publicly to facilitate ongoing work in understanding and extending the biological and clinical capabilities of CT-based foundation models.

In summary, our study demonstrates that vision-language foundation models for three-dimensional CT imaging can encode biologically and clinically meaningful variation that extends beyond the findings typically associated with CT indications. To our knowledge, Percival is the most comprehensively trained foundation model for CT imaging to date, enabling a systematic evaluation of the biological, phenotypic, and prognostic structure captured within a CT-VLM's latent space. By releasing Percival publicly, we aim to provide a broadly accessible resource for advancing future work in precision medicine and for deepening our understanding of how medical foundation models represent human biology and disease.

METHODS

Study population

The Penn Medicine BioBank (PMBB) comprises more than 300,000 consenting patients who receive care in the Penn Medicine healthcare system[42]. 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 over 51,000 PMBB participants. For validation, we included over 55,000 paired CT-report volumes from over 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[43]. 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 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) from the PyTorch Image Models (TIMM) library[44]. To process volumetric CT data, the vision transformer was adapted for three-dimensional imaging using three-dimensional patch embeddings, where each patch represents a 64 × 64 × 64 voxel segment that is projected into a latent embedding space using a 3D convolutional layer with kernel and stride equal to the patch size. From each CT volume, 32 regional patch embeddings and a single global patch embedding are extracted to ensure comprehensive spatial representation. Within the transformer architecture, each regional patch embedding serves as a token input to the encoder layers, where local contextual relationships are modeled through multi-head self-attention (MSA). The global patch embedding functions analogously to a [CLS] token and participates in the same attention operations, attending to and being attended by all regional patches at each layer. Through this bidirectional exchange, the global embedding progressively integrates fine-grained anatomical and

textural information from across the entire scan. The attention weights quantify the contribution of each regional patch to the global context, allowing the model to capture long-range dependencies between distant anatomical regions. 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[15,45,46]. The model was trained using a vectorized latent feature space of size 512, aligning imaging and text embeddings to optimize multimodal representation learning.

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[47,48]. 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 $([(x_{img}^1, x_{txt}^1), (x_{img}^2, x_{txt}^2), \dots, (x_{img}^b, x_{txt}^b)])$, the vision and language encoders (G_{img}, G_{txt}) compute corresponding latent feature representations $([(z_{img}^1, z_{txt}^1), (z_{img}^2, z_{txt}^2), \dots, (z_{img}^b, z_{txt}^b)])$. 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[49,50] 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 distributed data parallel (DDP) using two Nvidia A100 GPUs.

External validation: zero-shot classification

To assess out-of-domain generalization, we evaluated Percival's zero-shot disease classification performance on the CT-RATE external validation dataset, which included CT volumes from 1,304 patients and 18 distinct clinical indications which served as classification tasks[16]. Percival was benchmarked against two comparison models, Merlin and CT-CLIP[15,16]. Both Percival and Merlin were evaluated in a strict domain generalization setting, as neither model was fine-tuned on any portion of CT-RATE. In contrast, CT-CLIP was trained directly on the CT-RATE training split and therefore serves as an upper-bound reference for in-domain performance. Zero-shot classification was conducted following the same protocol as CT-CLIP, wherein the text encoder was provided with two standardized prompts: "[pathology] is present" and "[pathology] is not present." Pathology refers to the current CT indication. The resulting text feature embeddings were compared to image embeddings using cosine similarity, and the prompt yielding the higher similarity score was selected as the predicted disease label.

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), which together captured approximately half of the total variance. We focused on these leading components to provide a low-dimensional, interpretable summary of dominant axes of variation in the learned representation space and to enable stable statistical analysis across a wide range of clinical and laboratory traits. While disease-related signal may also be distributed across higher-order components, our objective was not to exhaustively characterize all latent dimensions, but rather to establish whether biologically meaningful structure is present among the principal modes of variation learned during pretraining. We first performed a univariate regression of each principal component with age (at the time of imaging), height (cm), weight (kg), systolic blood pressure (mmHg) and diastolic blood pressure (mmHg), as well as evaluating sex-differences within components using Cohen's D. 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 71 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 or 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.

Synthetic laboratory value estimation

For each laboratory measurement, we developed synthetic laboratory estimation models using linear probing of Percival's 512-dimensional latent feature space. Linear probing was implemented with ElasticNet regression via the glmnet package in R, using $\alpha = 1.0$ (Lasso). Predictor variables included Percival's 512-dimension feature space as well as

the difference between the date of imaging and the date of lab measurement. Models were trained and evaluated with an 80:20 train-validation split. Laboratory response variables were standardized with z-score normalization prior to model fitting to prevent scale-dependent effects and to decouple predicted values from assay-specific units. Model performance was quantified using the coefficient of determination (R^2) on the held-out validation set, enabling in silico estimation of physiological laboratory measurements directly from imaging-derived representations. During inference, the number of days between imaging and assay is set to 0, thus enabling synthetic estimation at the time of imaging.

Phenotypic associations

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-9/10CM codes that occurred within 60 days of image acquisition, whereas participants with earlier diagnosis dates were excluded from that analysis. PheWAS modeling was performed for each of the 10 PCs with adjustment for age and sex. Logistic regression models were then used to classify disease states. Phenome-wide significance was determined following FDR correction at the 5% significance threshold.

Lifespan Disease Burden Mapping

To assess global disease accumulation as a function of latent biological features, we quantified post-imaging disease burden for all participants using ICD-9/10CM codes recorded after the date were retained. ICD metadata were used to assign each diagnosis to one of 7 major clinical categories (e.g., circulatory, respiratory, endocrine/metabolic). Disease burden was summarized for each participant as the total number of unique ICD codes and the number of unique diagnosis codes within each ICD chapter. Participants' latent image embeddings were projected onto a 512-dimensional feature space and reduced using principal component analysis (PCA). Each principal component was z-score normalized and discretized into integer bins (-3 to +3). Average disease burden per-participant was computed within each bin and visualized as a heatmap to reveal global trends linking latent biological features to future disease accumulation.

Percival classification models

To perform disease-state classification, we developed logistic regression models for 378 disease states using the first ten PCs as predictors in the complete PMBB validation cohort (n=21,875). Consistent with the PheWAS design, ICD-9/10CM codes occurring within 60 days of imaging were used to define positive cases. Disease states with fewer than 50 positive cases were omitted from the analysis. To evaluate the predictive value of the latent feature space, 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 (AUC). 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 FDR correction applied for multiple comparisons and significance established at the 5% threshold[32]. All analyses were repeated separately in males and females to assess sex-specific model performance under the same framework.

Following disease classification, we performed targeted analyses of representative conditions spanning the lower and upper extremes of disease prevalence. For each selected disease, we examined the latent feature distributions of affected and unaffected participants using PCA derived from Percival's feature embeddings. Differences in feature distributions were visualized using density plots and quantified using Cohen's *d* effect sizes, providing a standardized measure of separation between groups. Effect sizes were grouped into small ($d<0.2$), medium ($d>0.5$) and large ($d>0.8$). These analyses allowed us to assess how disease prevalence influences the distinctiveness of latent feature representations and to identify patterns that may distinguish common from rare clinical phenotypes.

Percival prognostic risk modeling

In a similar manner, Cox proportional-hazards models were constructed using the first 10 latent feature space PCs, using ICD-9/10CM 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 with Bonferroni correction performed to adjust for multiple comparisons.

DATA AVAILABILITY

Access to Penn Medicine BioBank data is provided to investigators at the University of Pennsylvania. Percival's model architecture, 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

C.A.B is supported by NIH grant F31-HL182332. W.R.W. is supported by NIH grants P41-EB029460, R01-HL169378, R01-HL137984, UL1-TR001878, R21-EB036734, OT2-OD038048. J.A.C. is supported by NIH grants R01-HL 121510, R33-HL-146390, R01-HL153646, R01-AG058969, R01-HL104106, P01-HL094307, R03-HL146874, K24-AG070459 and R56-HL136730. J.G is supported by R01-EB031722, R01-HL133889. L.S. is supported by P30-AG073105, U01-AG088658 and U01-AG066833. MDR is supported by UL1-TR001878. E.E is supported by NIH grant OT2-OD038048.

DISCLOSURE

Dr. Chirinos is supported by NIH grants U01-TR003734, U01-TR003734-01S1, UO1-HL160277, R33-HL146390, 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. Rozenshtein, A., Findeiss, L.K., Wood, M.J., Shih, G. & Parikh, J.R. The U.S. Radiologist Workforce: AJR Expert Panel Narrative Review. *American Journal of Roentgenology* **224**, e2432085 (2025).
6. Dibble, E.H., Rubin, E. & Parikh, J.R. Workforce Shortage and Strategies for Mitigation: Results from the 2022 ACR/Radiology Business Management Association Workforce Survey. *Journal of the American College of Radiology* **22**, 573-576 (2025).
7. 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).
8. Ganeshan, D. *et al.* Burnout in Academic Radiologists in the United States. *Academic Radiology* **27**, 1274-1281 (2020).
9. 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).
10. Tushar, F.I. *et al.* Classification of Multiple Diseases on Body CT Scans Using Weakly Supervised Deep Learning. *Radiology: Artificial Intelligence* **4**, e210026 (2022).
11. 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).
12. Pickhardt, P.J. Value-added Opportunistic CT Screening: State of the Art. *Radiology* **303**, 241-254 (2022).
13. Wasserthal, J. *et al.* TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence* **5**, e230024 (2023).
14. Huang, J. *et al.* Efficiency and Quality of Generative AI–Assisted Radiograph Reporting. *JAMA Network Open* **8**, e2513921-e2513921 (2025).
15. Blankemeier, L. *et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography*, (2024).
16. Hamamci, I.E. *et al.* Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. (2024).
17. Bellemare, F., Jeanneret, A. & Couture, J. Sex Differences in Thoracic Dimensions and Configuration. *American Journal of Respiratory and Critical Care Medicine* **168**, 305-312 (2003).

18. Janssen, I., Heymsfield, S.B., Wang, Z. & Ross, R. Skeletal muscle mass and distribution in 468 men and women aged 18–88 yr. *Journal of Applied Physiology* **89**, 81-88 (2000).
19. Bredella, M.A. Sex Differences in Body Composition. in *Sex and Gender Factors Affecting Metabolic Homeostasis, Diabetes and Obesity* (ed. Mauvais-Jarvis, F.) 9-27 (Springer International Publishing, Cham, 2017).
20. 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).
21. 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).
22. 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).
23. Vagvala, S.H. & O'Connor, S.D. Imaging of abnormal liver function tests. *Clinical Liver Disease* **11**(2018).
24. Siddique, A. & Kowdley, K.V. Approach to a Patient with Elevated Serum Alkaline Phosphatase. *Clinics in Liver Disease* **16**, 199-229 (2012).
25. Selvin, E. *et al.* Glycated Hemoglobin, Diabetes, and Cardiovascular Risk in Nondiabetic Adults. *New England Journal of Medicine* **362**, 800-811 (2010).
26. Ference, B.A. *et al.* Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *European Heart Journal* **38**, 2459-2472 (2017).
27. Nordestgaard, B.G., Benn, M., Schnohr, P. & Tybjærg-Hansen, A. Nonfasting Triglycerides and Risk of Myocardial Infarction, Ischemic Heart Disease, and Death in Men and Women. *JAMA* **298**, 299-308 (2007).
28. Assmann, G., Schulte, H., von Eckardstein, A. & Huang, Y. High-density lipoprotein cholesterol as a predictor of coronary heart disease risk. The PROCAM experience and pathophysiological implications for reverse cholesterol transport. *Atherosclerosis* **124**, S11-S20 (1996).
29. Hilton, C., Sabaratnam, R., Drakesmith, H. & Karpe, F. Iron, glucose and fat metabolism and obesity: an intertwined relationship. *International Journal of Obesity* **47**, 554-563 (2023).
30. Sproston, N.R. & Ashworth, J.J. Role of C-Reactive Protein at Sites of Inflammation and Infection. *Frontiers in Immunology* **Volume 9 - 2018**(2018).
31. 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).
32. 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).
33. 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).

34. 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).
35. You, K. *et al.* CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training. *ArXiv* **abs/2310.13292**(2023).
36. Kim, C. *et al.* Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine* **30**, 1154-1165 (2024).
37. Chen, Z. *et al.* A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation. (2024).
38. Zhang, X. *et al.* Tubular secretion of creatinine and kidney function: an observational study. *BMC Nephrology* **21**, 108 (2020).
39. Garimella, P.S., Tighiouart, H., Sarnak, M.J., Levey, A.S. & Ix, J.H. Tubular Secretion of Creatinine and Risk of Kidney Failure: The Modification of Diet in Renal Disease (MDRD) Study. *American Journal of Kidney Diseases* **77**, 992-994 (2021).
40. 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).
41. 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).
42. 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).
43. 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).
44. Touvron, H. *et al.* Training data-efficient image transformers & distillation through attention. in *International Conference on Machine Learning* (2020).
45. 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).
46. 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).
47. Oord, A.v.d., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. *ArXiv* **abs/1807.03748**(2018).
48. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. in *International Conference on Machine Learning* (2021).
49. Kingma, D.P. & Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* **abs/1412.6980**(2014).
50. Loshchilov, I. & Hutter, F. Fixing Weight Decay Regularization in Adam. *ArXiv* **abs/1711.05101**(2017).
51. NIAID VIsual & Medical Arts. (NIAID NIH BIOART, 2024).

Figure legends

Figure 1. Overview of data, model design, and analyses. (**A**) An overview of the Penn Medicine Biobank (PMBB) computed-tomographic (CT) imaging data used in this study, and the train-validation split. (**B**) An illustration of Percival's vision encoder architecture which leverages an augmented vision transformer (ViT) to support three-dimensional patch embeddings. (**C**) An illustration of Percival's language encoder architecture which leverages the pretrained clinical Longformer architecture. (**D**) Volumetric CT imaging data and radiology report text data are aligned in a 512-dimension latent feature space using a symmetric contrastive (INFO NCE) loss function. (**E**) Evaluation of Percival's biological alignment. (**F**) Contextualization of Percival's latent feature space alignment with disease phenotypes and lifespan disease burden. (**G**) Evaluation of classification and prognostic risk stratification models for inferring disease state and progression.

Figure 2. Biological associations of Percival's latent space. (**A**) Association between age and latent PC1 shown using linear regression. (**B**) Density plots of latent PCs 1-10 stratified by sex. Color intensity reflects the magnitude of Cohen's D effect size for male-female separation. (**C**) Association between weight and latent PC5 shown using linear regression. (**D**) Half-half discovery–replication laboratory-wide association study (LabWAS) for latent PCs 1-5. An "X" indicates associations significant after 5% FDR correction in both cohorts; "I" and "O" indicate significance in the discovery and replication cohorts, respectively. (**E**) Training and validation Pearson R correlation coefficients for synthetic laboratory value prediction models derived from Percival's latent features.

Figure 3. Disease-enrichment patterns and lifespan disease burden. (**A**) Circular Manhattan plot depicting phenome-wide associations between Percival's latent PCs and clinical diagnoses across 7 ICD chapters. The red dashed line marks the phenome-wide significance threshold after 5% FDR correction. Upward triangles indicate positive associations with disease prevalence; downward triangles indicate negative associations. (**B**) Lifespan disease-burden mapping for z-score–normalized latent PCs, showing the average number of chapter-specific diagnoses accumulated across participants as a

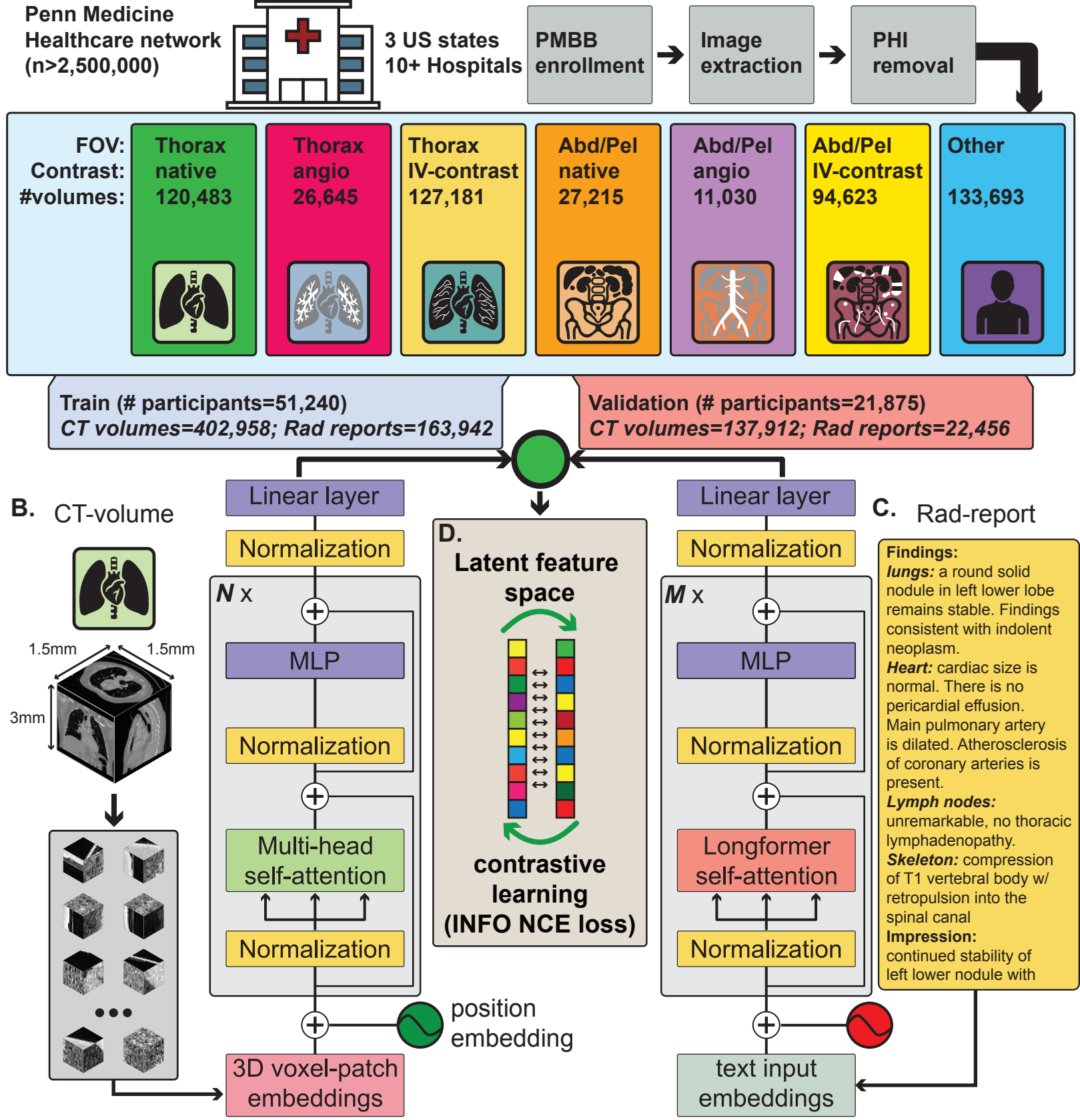
function of PC value. (**C**) Bar plot illustration of circulatory disease burden for component 3. (**D**) Bar plot illustration of endocrine/metabolic disease burden for component 5.

Figure 4. Temporal disease classification. Using ICD-9/10CM codes within 60 days of imaging we evaluated Percival's ability to perform temporal disease classification. (**A**) Density plot of the number of positive disease cases (log-transformed) for all classification tasks. Colors denote cumulative proportions of ≤50%, ≤75%, ≤90%, and >90% of diseases by prevalence. (**B**) Relationship between five-fold cross-validated AUC performance and the number of positive disease cases for each outcome. (**C**) Five-fold cross-validated AUC and 95% confidence intervals for selected disease states across 7 ICD chapters. Human anatomical art provided by NIH BIOART[51].

Figure 5. Disease classification across the disease-prevalence spectrum. Three representative outcomes were examined to assess how Percival's latent features distinguish affected from unaffected participants across low-, moderate-, and high-prevalence conditions. Each disease is shown in three panels: (**1**) a five-fold cross-validated area under the receiver operating characteristic curve (AUC) comparing Percival to a demographic baseline model, (**2**) component-level separability visualized by Cohen's D effect sizes plotted against mean differences, and (**3**) density plots of principal component scores stratified by outcome, with color indicating the magnitude of Cohen's D. (**A–C**) Respiratory failure (low prevalence; n=50) demonstrated strong discriminative performance (AUC=0.81) with prominent separation on higher-order components. (**D–F**) Aortic aneurysm (moderate prevalence; n=339) achieved an AUC of 0.79 with its largest effects on PC1 and PC8. (**G–I**) chronic kidney disease (high prevalence; n=657) showed more diffuse separability with modest effect sizes distributed across multiple components.

Figure 6. Prognostic utility of Percival's latent feature space using Cox proportional hazards modeling. The first ten principal components of Percival's latent representations were used as inputs to Cox models, with adjustment for age and sex. Linear predictor values from the fitted models were divided into tertiles to define low-, intermediate-, and high-risk groups. Kaplan-Meier curves illustrate risk stratification across four

representative disease states: (A) respiratory failure, (B) chronic pulmonary heart disease, (C) cachexia, and (D) acute renal failure. Human anatomical art provided by NIH BIOART[51].



E. Biological alignment

- Associations with clinical and physiological traits
- Correspondence with laboratory biomarkers
- Inferring synthetic laboratory values

F. Structure of latent space

- Characterizing lifespan and disease burden trajectories
- Mapping phenotypic associations
- Biological alignment across disease prevalence

G. Disease modeling & risk

- Linear probing of 378 disease states
- Prognostic modeling of 648 disease states using Cox proportional hazards
- Stratification of longitudinal risk

Figure 1.

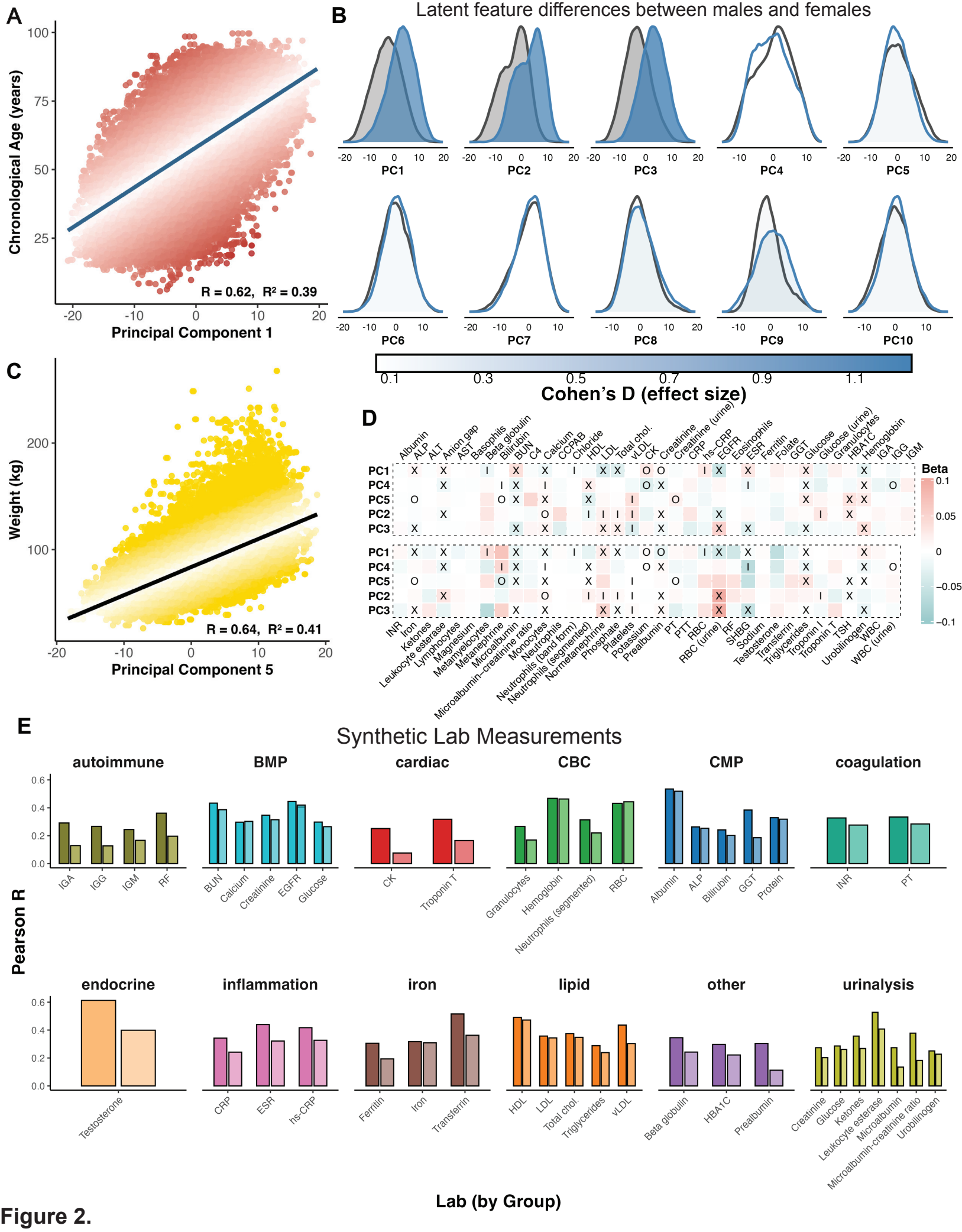

A
Chronological Age (years)
Principal Component 1
R = 0.62, R² = 0.39
B
Latent feature differences between males and females
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
Cohen's D (effect size)
C
Weight (kg)
Principal Component 5
R = 0.64, R² = 0.41
D
Beta
E
Synthetic Lab Measurements
autoimmune
BMP
cardiac
CBC
CMP
coagulation
endocrine
inflammation
iron
lipid
other
urinalysis
Pearson R
Lab (by Group)

Figure 2.

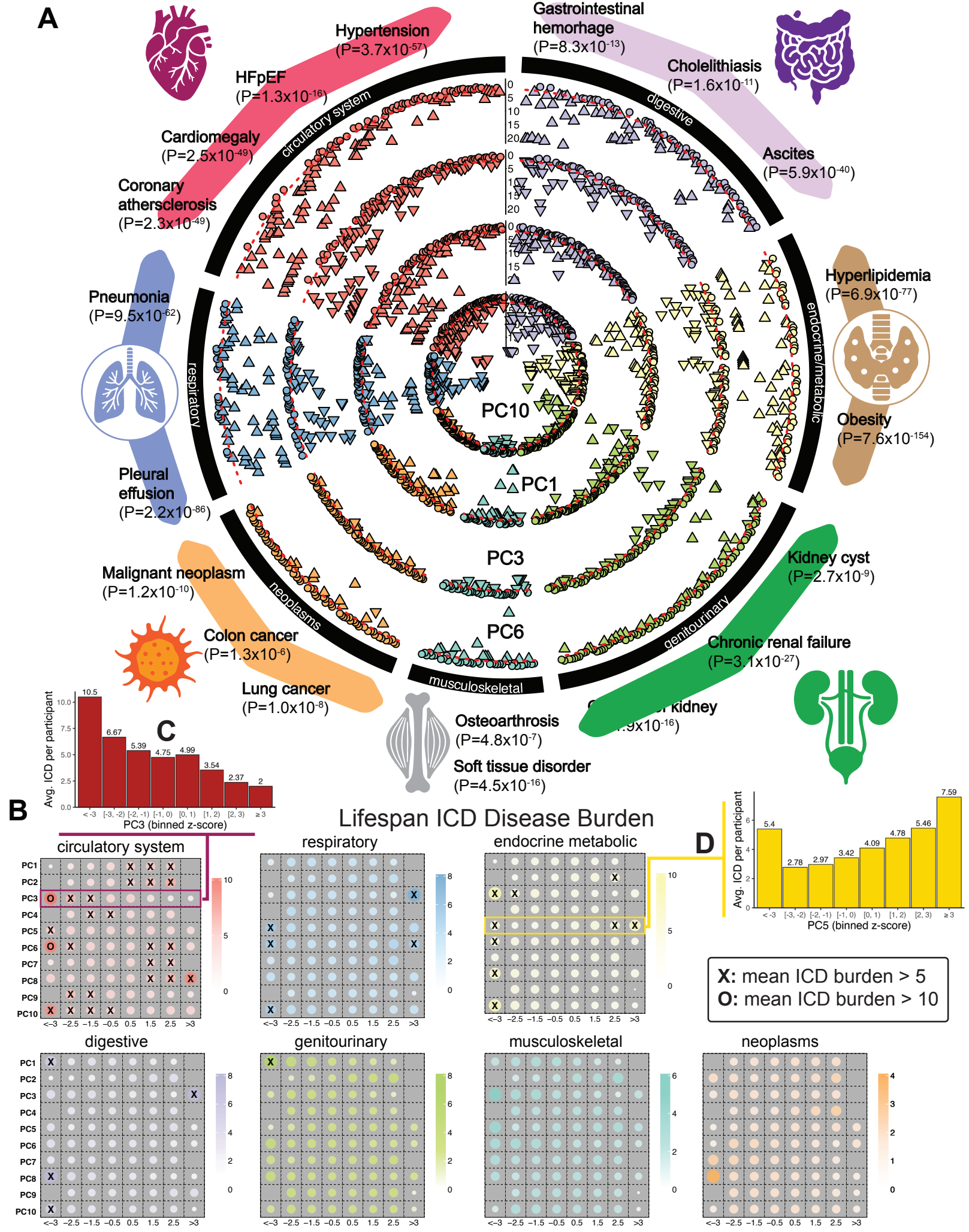

A
Hypertension
(P=3.7x10-57)
HFpEF
(P=1.3x10-16)
Cardiomegaly
(P=2.5x10-49)
Coronary
atherosclerosis
(P=2.3x10-49)
Pneumonia
(P=9.5x10-62)
Pleural
effusion
(P=2.2x10-86)
Malignant neoplasm
(P=1.2x10-10)
Colon cancer
(P=1.3x10-6)
Lung cancer
(P=1.0x10-8)
Osteoarthrosis
(P=4.8x10-7)
Soft tissue disorder
(P=4.5x10-16)
Kidney cyst
(P=2.7x10-9)
Chronic renal failure
(P=3.1x10-27)
Hyperlipidemia
(P=6.9x10-77)
Obesity
(P=7.6x10-154)
Ascites
(P=5.9x10-40)
Cholelithiasis
(P=1.6x10-11)
Gastrointestinal
hemorhage
(P=8.3x10-13)
circulatory system
digestive
endocrine/metabolic
genitourinary
musculoskeletal
neoplasms
respiratory
PC10
PC1
PC3
PC6
C
Avg. ICD per participant
PC3 (binned z-score)
B
Lifespan ICD Disease Burden
circulatory system
respiratory
endocrine metabolic
D
Avg. ICD per participant
PC5 (binned z-score)
X: mean ICD burden > 5
O: mean ICD burden > 10
digestive
genitourinary
musculoskeletal
neoplasms

Figure 3.

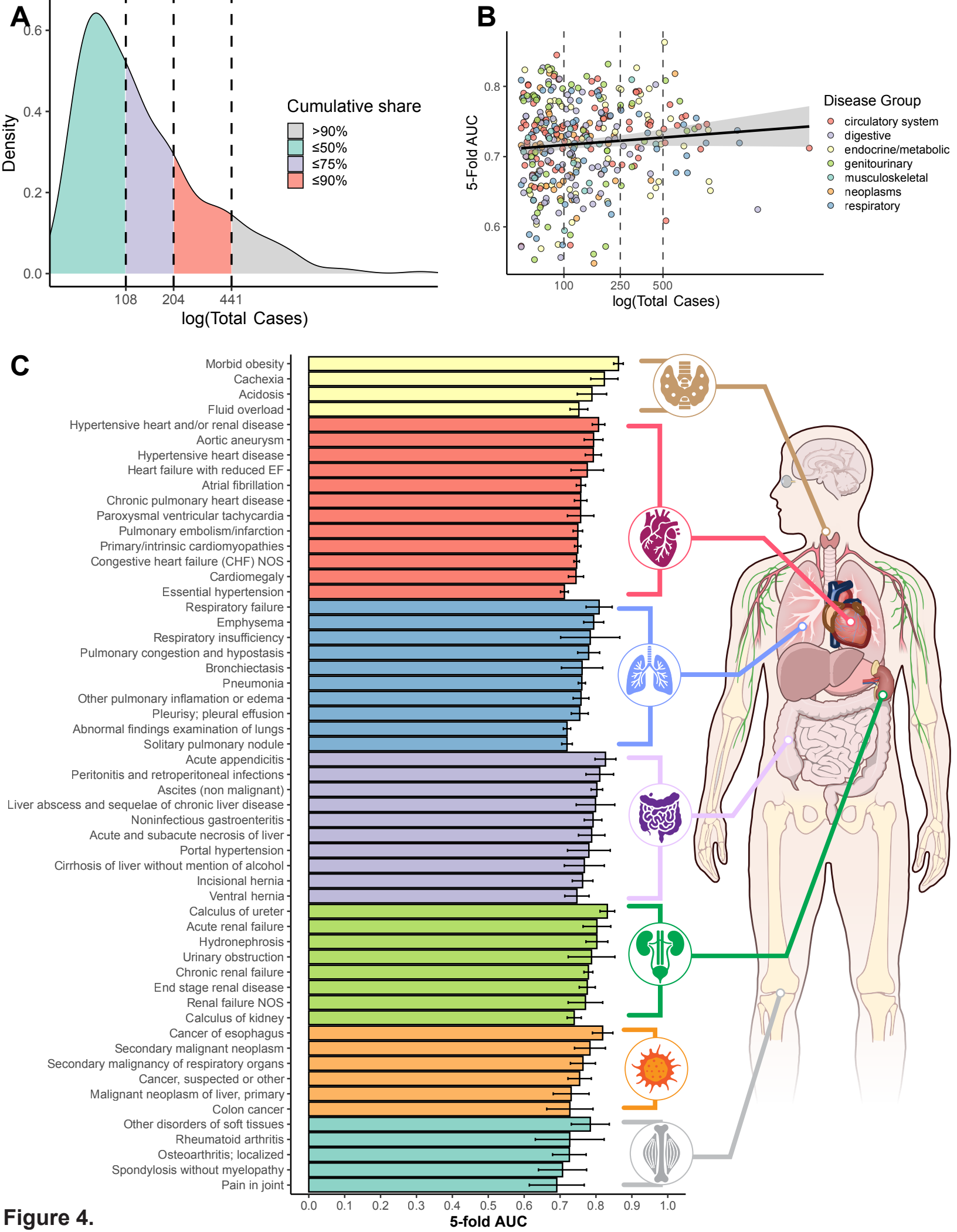

A
Density
0.0
0.2
0.4
0.6
108
204
441
log(Total Cases)
Cumulative share
>90%
≤50%
≤75%
≤90%
B
5-Fold AUC
0.6
0.7
0.8
100
250
500
log(Total Cases)
Disease Group
circulatory system
digestive
endocrine/metabolic
genitourinary
musculoskeletal
neoplasms
respiratory
C
Morbid obesity
Cachexia
Acidosis
Fluid overload
Hypertensive heart and/or renal disease
Aortic aneurysm
Hypertensive heart disease
Heart failure with reduced EF
Atrial fibrillation
Chronic pulmonary heart disease
Paroxysmal ventricular tachycardia
Pulmonary embolism/infarction
Primary/intrinsic cardiomyopathies
Congestive heart failure (CHF) NOS
Cardiomegaly
Essential hypertension
Respiratory failure
Emphysema
Respiratory insufficiency
Pulmonary congestion and hypostasis
Bronchiectasis
Pneumonia
Other pulmonary inflamation or edema
Pleurisy; pleural effusion
Abnormal findings examination of lungs
Solitary pulmonary nodule
Acute appendicitis
Peritonitis and retroperitoneal infections
Ascites (non malignant)
Liver abscess and sequelae of chronic liver disease
Noninfectious gastroenteritis
Acute and subacute necrosis of liver
Portal hypertension
Cirrhosis of liver without mention of alcohol
Incisional hernia
Ventral hernia
Calculus of ureter
Acute renal failure
Hydronephrosis
Urinary obstruction
Chronic renal failure
End stage renal disease
Renal failure NOS
Calculus of kidney
Cancer of esophagus
Secondary malignant neoplasm
Secondary malignancy of respiratory organs
Cancer, suspected or other
Malignant neoplasm of liver, primary
Colon cancer
Other disorders of soft tissues
Rheumatoid arthritis
Osteoarthritis; localized
Spondylosis without myelopathy
Pain in joint
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
5-fold AUC

Figure 4.

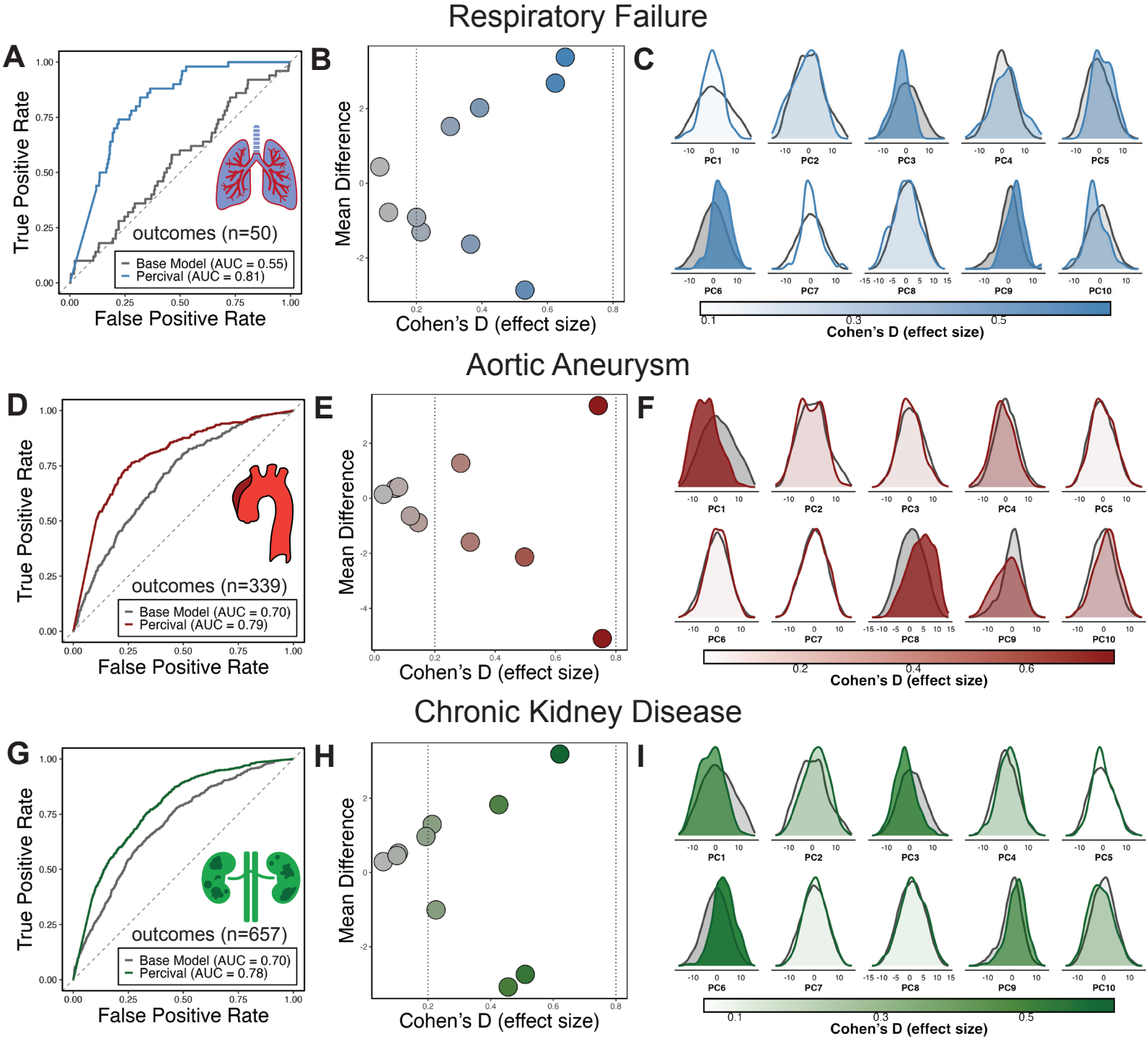

Respiratory Failure
A
True Positive Rate
False Positive Rate
outcomes (n=50)
Base Model (AUC = 0.55)
Percival (AUC = 0.81)
B
Mean Difference
Cohen's D (effect size)
C
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
Cohen's D (effect size)
Aortic Aneurysm
D
True Positive Rate
False Positive Rate
outcomes (n=339)
Base Model (AUC = 0.70)
Percival (AUC = 0.79)
E
Mean Difference
Cohen's D (effect size)
F
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
Cohen's D (effect size)
Chronic Kidney Disease
G
True Positive Rate
False Positive Rate
outcomes (n=657)
Base Model (AUC = 0.70)
Percival (AUC = 0.78)
H
Mean Difference
Cohen's D (effect size)
I
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
Cohen's D (effect size)

Figure 5.

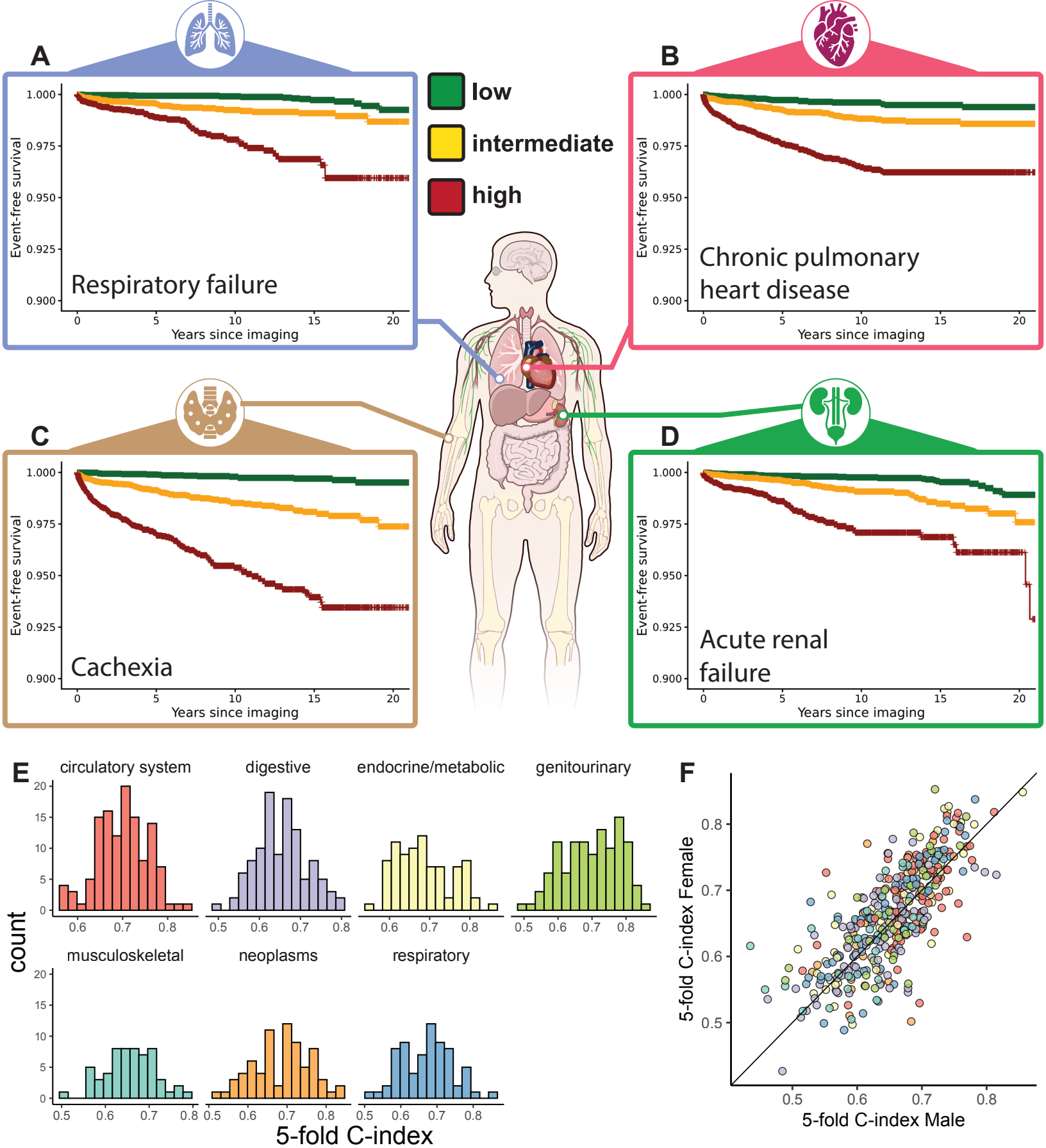

A
Respiratory failure
B
Chronic pulmonary heart disease
C
Cachexia
D
Acute renal failure
low
intermediate
high
Event-free survival
Years since imaging
E
circulatory system
digestive
endocrine/metabolic
genitourinary
musculoskeletal
neoplasms
respiratory
count
5-fold C-index
F
5-fold C-index Female
5-fold C-index Male

Figure 6.

Supplemental Figures

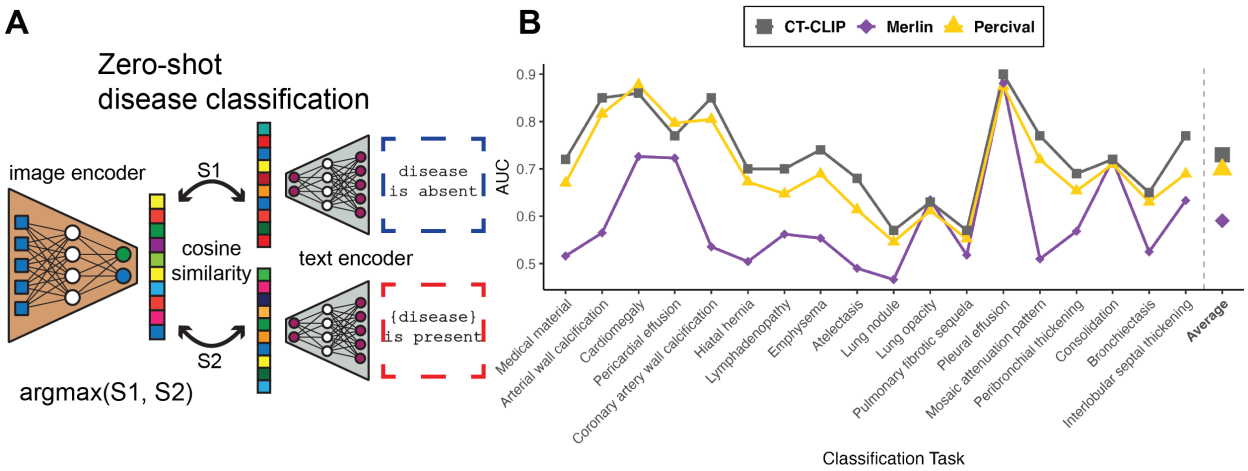


Figure S1. (**A**) Schematic of zero-shot disease classification. Text prompts ("[finding] is present" / "[finding] is not present") are encoded and compared with CT embeddings via cosine similarity; the higher similarity determines the predicted label. (**B**) Zero-shot classification performance on the CT-RATE dataset for Percival, Merlin, and CT-CLIP. Percival and Merlin were evaluated strictly out-of-domain, whereas CT-CLIP, trained on the CT-RATE training split, serves as an upper-bound in-domain reference. CT-CLIP performance values are from the original publication[1].

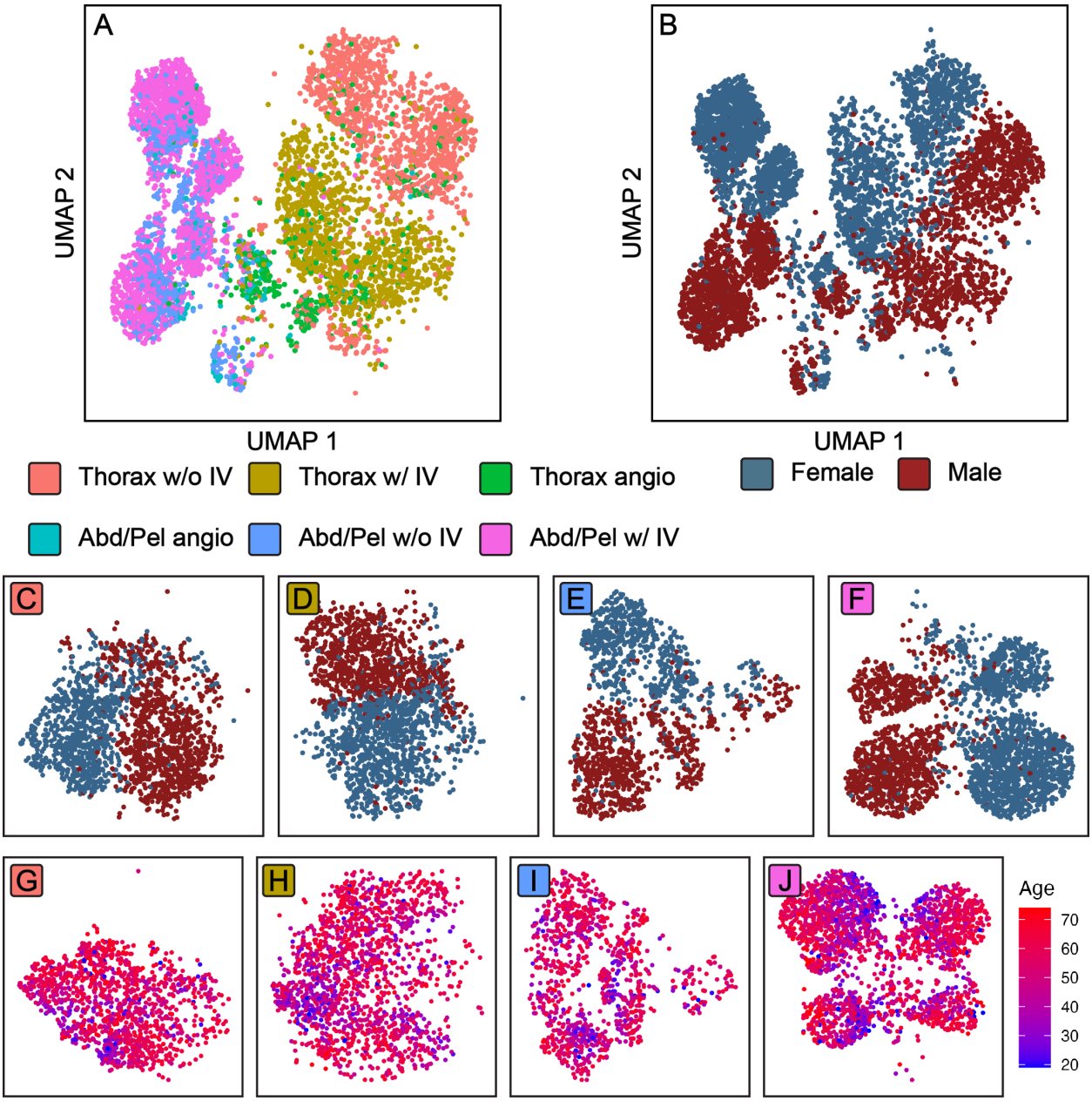


Figure S2. UMAP dimensionality reduction of Percival's vision encoder on validation data. UMAP visualizations of image embeddings from the Percival model. (**A**) UMAP plot color-coded for CT fields of view and contrast method. (**B**) UMAP plot color-coded for sex. (**C**) UMAP plot of thoracic CT volumes without IV contrast, color-coded for sex. (**D**) UMAP plot of thoracic CT volumes with IV contrast, color-coded for sex. (**E**) UMAP plot of abdominal and pelvic CT volumes without IV contrast, color-coded for sex. (**F**) UMAP plot of abdominal and pelvic CT volumes with IV contrast, color-coded for sex. (**G**) UMAP plot of thoracic CT volumes without IV contrast, color-coded for age. (**H**) UMAP

plot of thoracic CT volumes with IV contrast, color-coded for age. (**I**) UMAP plot of abdominal and pelvic CT volumes without IV contrast, color-coded for age. (**J**) UMAP plot of abdominal and pelvic CT volumes with IV contrast, color-coded for age.

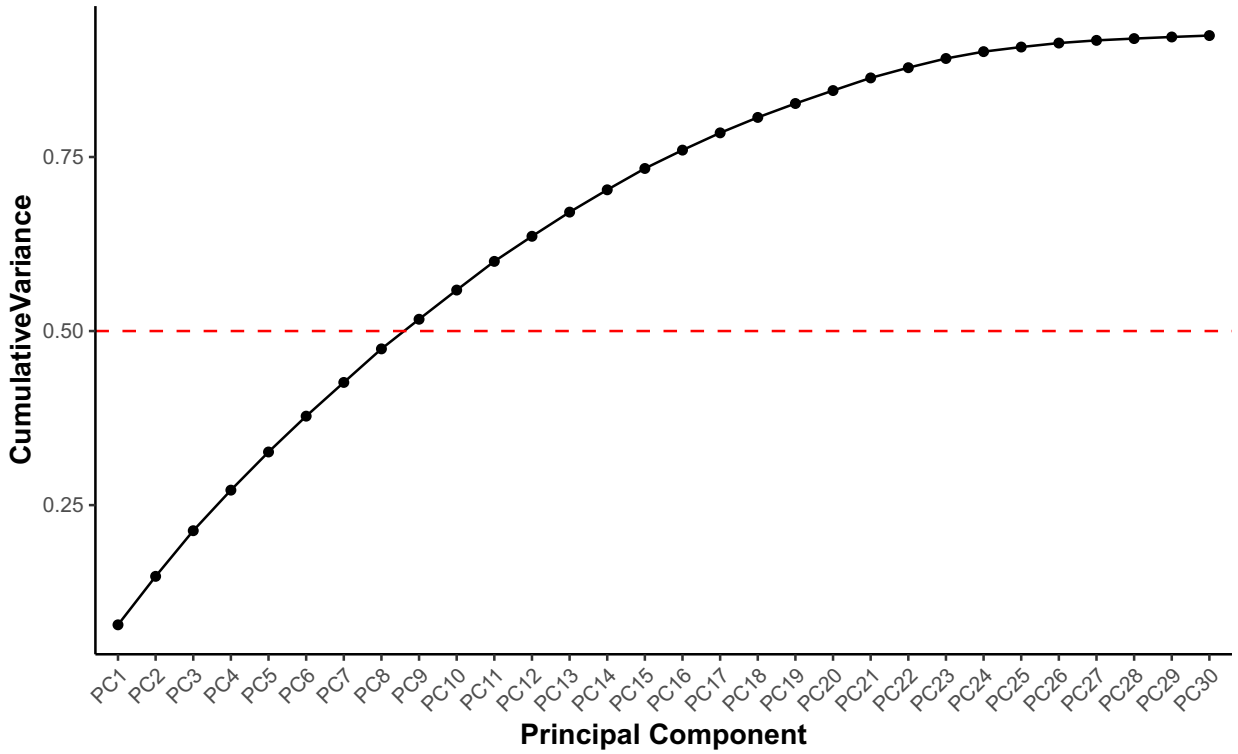


Figure S3. Cumulative variance explained by the first 30 principal components derived from Percival's latent feature embeddings. The x-axis represents individual principal components (PCs), and the y-axis indicates the cumulative proportion of variance explained. A dashed red line marks the 50% threshold.

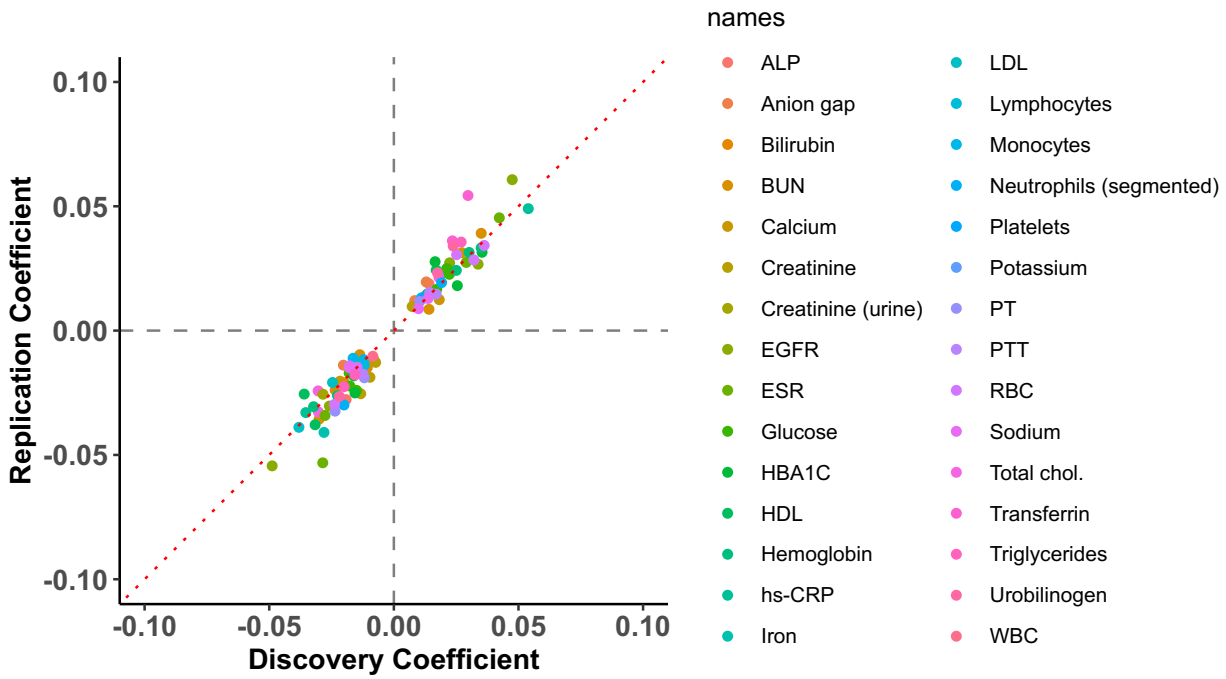


Figure S4. Discovery-replication concordance of associations for laboratory measurements. Significant associations following FDR-adjusted multiple comparisons between principal components derived from Percival's latent feature space and clinical laboratory measurements in the discovery (average n=3,265) and replication (average n=3,268) cohorts.

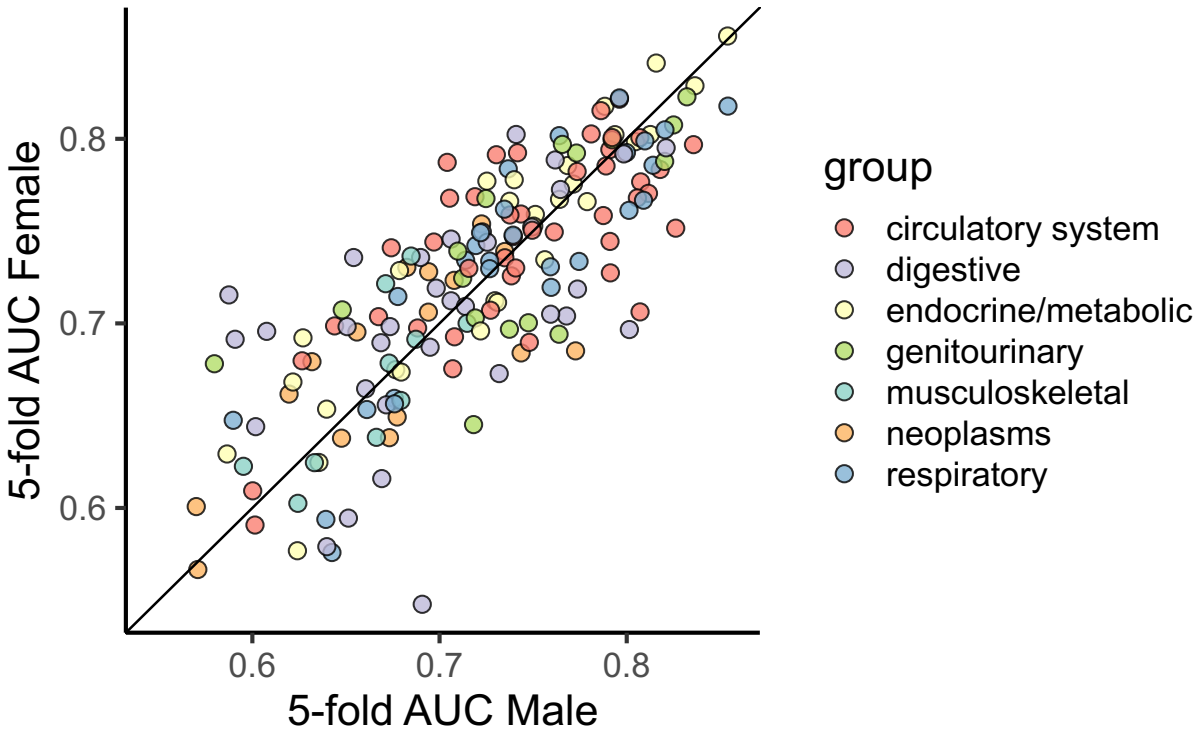


Figure S5. Sex-specific concordance of phenome-wide associations across 7 ICD chapters. Area under the receiver operating characteristic curve (AUC) for sex-specific models for males and females. Color coding indicates ICD-chapter.

References

1. Hamamci, I.E. *et al.* Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. (2024).

Supplemental Note

Penn Medicine BioBank Banner Author List and Contribution Statements

PMBB Leadership Team

Daniel J. Rader, M.D., Marylyn D. Ritchie, Ph.D.

Contribution: All authors contributed to securing funding, study design and oversight. All authors reviewed the final version of the manuscript.

Patient Recruitment and Regulatory Oversight

JoEllen Weaver, Nawar Naseer, Ph.D., M.P.H., Giorgio Sirugo, M.D., P.h.D., Afiya Poindexter, Yi-An Ko, Ph.D., Kyle P. Nerz

Contributions: JW manages patient recruitment and regulatory oversight of study. NN manages participant engagement, assists with regulatory oversight, and researcher access. GS assists with researcher access. AP, YK, KPN perform recruitment and enrollment of study participants.

Lab Operations

JoEllen Weaver, Meghan Livingstone, Fred Vadivieso, Stephanie DerOhannessian, Teo Tran, Julia Stephanowski, Salma Santos, Ned Haubein, P.h.D., Joseph Dunn

Contribution: JW, ML, FV, SD conduct oversight of lab operations. ML, FV, AK, SD, TT, JS, SS perform sample processing. NH, JD are responsible for sample tracking and the laboratory information management system.

Clinical Informatics

Anurag Verma, Ph.D., Colleen Morse Kripke, M.S. DPT, MSA, Marjorie Risman, M.S., Renae Judy, B.S., Colin Wollack, M.S.

Contribution: All authors contributed to the development and validation of clinical phenotypes used to identify study subjects and (when applicable) controls.

Genome Informatics

Anurag Verma Ph.D., Shefali S. Verma, Ph.D., Scott Damrauer, M.D., Yuki Bradford, M.S., Scott Dudek, M.S., Theodore Drivas, M.D., Ph.D.,

Contribution: AV, SSV, and SD are responsible for the analysis, design, and infrastructure needed to quality control genotype and exome data. YB performs the analysis. TD and AV provides variant and gene annotations and their functional interpretation of variants.