
Supplementary information

Thymic health consequences in adults

In the format provided by the
authors and unedited

Supplementary Information

SUPPLEMENTARY METHODS

Supplementary Methods S1 Deep Learning Pipeline Development	2
1.1. Data	2
1.2. Center of Mass (CoM) Localization	2
1.3. Quantification of the Thymus	2
1.4. Architecture selection ablations	2
1.5. Evaluation of the compared methods	4
1.6. Inter-rater analyses and quality control	4
1.7. Software	4
Supplementary Methods S2 Deep Learning Pipeline Evaluation	6
2.1. Evaluation of deep learning architectures and learning algorithms	6
2.2. Quality control	6
2.2.1. Inter-rater analyses (human vs human)	6
2.2.2. Validation of the automated CoM localizer	6
2.2.3. Quantification model robustness	7
2.3. AI-generated feature interpretability	7
2.4. Pipeline performance	8

SUPPLEMENTARY FIGURES

Figure 1 Development of the deep learning system	10
Figure 2 Association of age and body mass index (BMI) with thymic health	11
Figure 3 Association of thymic health with health outcomes in NLST using 3-year binned age stratification	12
Figure 4 Association of thymic health with long-term risk of lung cancer and lung cancer-specific mortality according to smoking status	13
Figure 5 Associations of thymic health with metabolism according to sex	14
Figure 6 Distribution of thymic bed sizes along each spatial axis	15
Figure 7 Representative computed tomography (CT) images of manually annotated thymic extreme representations	16
Figure 8 Evaluation of the deep learning system	17
Figure 9 Spatial distributions of pairwise comparisons between CoM annotations per axis	18
Figure 10 Distribution of thymus sizes along each spatial axis	19
Figure 11 Representative examples of thymic segmentations from the auto-localization model	20
Figure 12 Evaluation of the automated localizer model	21
Figure 13 Feature explainability of thymic health on FHS and NLST cohorts	22
Figure 14 Study population inclusion	23
Figure 15 Sensitivity analysis of thymic health according to independent manual thymus grading	24
Figure 16 Association of thymic health with thymic bed volume	25
Figure 17 Association of thymic bed volume with long-term mortality	26

SUPPLEMENTARY TABLES

Table 1 Distribution of diagnoses for malignant neoplasm-specific death	28
Table 2 Distribution of diagnoses for cardiovascular-specific death	29
Table 3 Distribution of diagnoses for pulmonary disease-specific death	30
Table 4 Distribution of diagnoses for endocrine, nutritional, and metabolic disease-specific death	31
Table 5 Distribution of diagnoses for digestive disease-specific death	32
Table 6 Fried frailty phenotype analysis in Framingham Heart Study participants	33
Table 7 OLINK Inflammation panel analysis in Framingham Heart Study participants	34
Table 8 Chronic inflammation analysis in Framingham Heart Study participants	35
Table 9 Data collections and respective link	36
Table 10 Details of hyperparameters used during the self-supervised pre-training process	37
Table 11 Confusion matrix for the inter-observer agreement between two expert radiologists	38
Table 12 Summary statistics of inter-rater CoM distances	39
Table 13 Summary statistics of raw CoM coordinates	40

SUPPLEMENTARY REFERENCES |

Supplementary Methods S1 | Deep Learning Pipeline Development

We developed an end-to-end deep learning system that automatically quantifies an individual's thymic characteristics from a computed tomography (CT) scan as a proxy for thymic functionality, and we termed this measure thymic health (**Supplementary Figure 1**). The deep learning system comprises a pipeline that first localizes the thymic bed and then quantifies the thymus. A thymic bed segmentation model using a U-Net-based architecture was trained to segment the thymic bed, from which a Center of Mass (CoM) is calculated to define the coordinates of the central localization of the thymus. This CoM serves as a reference to extract a 3D crop around it to be used for the quantification of the thymus. The thymic health quantification model was developed using, first, a self-supervised contrastive learning (SSL) task-agnostic pre-training followed by a fine-tuning of a shallow classifier on the SSL representations. During training of the quantification model, manually annotated CoMs were used to minimize noise. The end-to-end application of the system, however, uses automated CoM and was validated in presumed healthy individuals of larger cohorts (National Lung Screening Trial [NLST] and Framingham Heart Study [FHS]).

1.1. Data

For the localization algorithm, our training set was composed of 2,641 annotated CT scans covering the thoracic region, and for the pre-training of the quantification model, we collected a training dataset that included 5,674 CT scans covering the thoracic region. A subset of this dataset, with 1,637 CT scans, was used for the outcome-based fine-tuning. The CT scans were resampled to a 1mm isotropic spacing; then, the Hounsfield Units (HU) intensities were clipped between -1,024 and 3,072 HU and normalized to a range of 0 to 1 using min-max normalization. The DOIs for the data collections used to train the models can be found in **Supplementary Table 9**.

1.2. Center of Mass (CoM) Localization

For the localization step, a thymic bed segmentation U-Net⁴¹ model was trained in a supervised fashion on the training dataset. Our training approach itself leveraged the use of an active-learning framework⁴² in which we had an expert reader providing human-in-the-loop feedback throughout the training process. During inference, we predicted segmentation masks for scans across all datasets using this model and computed the respective center of mass for each scan through averaging coordinates.

1.3. Quantification of the Thymus

We developed a quantification process to automatically assess thymic health from CT scans. After performing comprehensive ablations across different deep-learning approaches (see **section 1.4. Architecture selection ablations**), we selected our final method based on performance and ease of implementation.

The quantification process works by pre-training a deep-learning model to extract meaningful features from thymic regions in CT scans. The model learns to capture patterns and characteristics of the thymus without relying on manual labels during training. During inference, the system generates high-dimensional feature representations that capture abstract properties of the thymic tissue, capturing broader compositional and morphological signals that extend beyond volume. These learned features are then used to fine-tune a prediction model (shallow classifier) that can automatically rank thymic health from low to high, validated against expert annotations.

This automated approach enables consistent and scalable assessment of thymic health across large datasets, reducing the need for manual expert evaluation while maintaining clinical accuracy.

1.4. Architecture selection ablations

We conducted a comprehensive evaluation of self-supervised learning and supervised methods to identify optimal architectures for thymic imaging analysis. The distinct approaches implemented and compared using self-supervised pre-training were: DINOv2⁴³, Masked Autoencoder (MAE⁴⁴), Nearest-Neighbor Contrastive Learning of Visual Representations (NNCLR⁴⁵), and Swapping Assignments between Views (SwAV⁴⁶). Supervised approaches using Vision Transformers (ViT) and ResNet-50 were also implemented.

DINOv2: The DINOv2 framework employed a 3D Vision Transformer with ViT-S/4 architecture configuration, featuring 8 transformer layers, 768 hidden dimensions, 3,072 MLP dimensions, and 12

attention heads. The model processed $48 \times 48 \times 48$ voxel patches with $6 \times 6 \times 6$ patch tokenization. Training utilized a batch size of 48 per device across 3 GPUs for 500 epochs with mixed precision (16-bit). The optimization strategy employed a base learning rate of 1×10^{-5} , layer decay of 0.9, gradient clipping at 3.0, and weight decay of 0.04. Teacher-student distillation parameters included temperature warmup over 30 epochs with minimum and maximum temperatures of 0.005, and projection dimensionality of 16,384. The framework incorporated iBOT methodology with separate prediction heads and applied domain-specific 3D augmentations optimized for medical imaging data.

MAE: The MAE implementation utilized a 3D encoder-decoder architecture with $48 \times 48 \times 48$ input resolution and $6 \times 6 \times 6$ patch tokenization. The encoder comprised 12 transformer layers with 768 embedding dimensions, while the decoder employed 4 layers for reconstruction. Training was conducted with a 75% masking ratio, optimizing reconstruction loss through pixel-level supervision of masked patches. The framework processed thymic patches through specialized 3D augmentations adapted for volumetric medical data.

SwAV: The SwAV implementation employed a contrastive learning framework based on clustering assignments between multiple augmented views of thymic patches. The architecture utilized a 3D ResNet-50 backbone with $2 \times$ widening factor, single input channel, and temporal stride 2, configured without final classification layers to serve as a feature encoder. The model incorporated a projection head with 4,096 feature dimensions mapped to 128-dimensional embeddings for clustering operations. The SwAV methodology implemented online clustering with 300 learnable prototypes, maintaining 4 queue structures, each with a length equivalent to 5 times the effective batch size (640 samples for batch size 128). Prototype assignment queues were activated after epoch 15, following the original SwAV protocol, with prototypes frozen for 1 step during each assignment update. The clustering loss employed the Sinkhorn-Knopp algorithm for optimal transport assignment with temperature parameter 0.1 and epsilon regularization 0.03. Training was conducted using stochastic gradient descent optimization with dynamically scaled learning rates based on batch size ($0.6 \times \text{batch_size}/256$), momentum 0.9, and weight decay 1×10^{-6} . Learning rate scheduling followed by cosine annealing with minimum rate $0.0006 \times \text{batch_size}/256$. Based on the thymus anatomical size distributions (see **Supplementary Figure 6**), the input to the self-supervised framework consists of thymic bed regions in scans obtained by cropping a region of 50 mm x 50 mm x 75 mm around the thymic bed CoM along the x, y, and z directions, respectively. The framework processed thymic patches at $75 \times 50 \times 50$ resolution through multiple augmented views generated via replicated transformations, including random resized cropping (scale 0.3-1.0), random affine transformations (rotation $\pm 10^\circ$, shear 0.1, zero padding), random histogram shifting, random Gaussian smoothing, each applied with 50% probability, and spatial padding to ensure the input shape of $75 \times 50 \times 50$.

NNCLR: The NNCLR implementation used a 3D ResNet50 backbone, with a widening factor of $2 \times$, single-input channel, and temporal stride of 2. Similarly to SwAV, the model's 4,096 features latent space is also mapped to a 128-dimensional embedding space for the clustering operations. The training loss was NNCLR_NegativeMinedLoss with a temperature of 0.1, optimized using Layer-wise Adaptive Rate Scaling (LARS) with dynamic learning rate updates based on batch size ($0.3 \times \text{batch_size}/256$), weight decay of 1×10^{-6} , and 0.9 of momentum, scheduled employing a cosine annealing strategy. The transforms used to pre-process the input patches included random resized cropping (scale 0.5-1.2), random affine transforms (rotation $\pm 3.5^\circ$, shear 0.1, zero-padding), histogram shifting, and Gaussian smoothing, each with 50% probability of usage, and finally spatial padding to ensure the $75 \times 50 \times 50$ shape.

ViT (Supervised): The supervised ViT implementation employed a 12-layer 3D Vision Transformer with 768 hidden dimensions and 3,072 MLP dimensions, containing approximately 114 million trainable parameters. Input patches were processed at $48 \times 48 \times 48$ resolution with $8 \times 8 \times 8$ patch size. Training utilized binary cross-entropy loss with logits for thymic classification, optimized via AdamW with learning rate 1×10^{-4} , weight decay 1×10^{-6} , and cosine annealing scheduling. Data augmentation included random Gaussian smoothing, spatial flipping, center cropping, spatial padding, affine transformations (rotation, shear, translation), and histogram shifting, all applied with 50% probability.

ResNet-50 (Supervised): The ResNet-50 baseline implemented a 3D convolutional architecture with approximately 184 million trainable parameters, configured with a widening factor of 2, single input channel, and temporal stride of 2 in the initial convolution. The model processed thymic regions at $75 \times 50 \times 50$ resolution through identical augmentation strategies as the supervised ViT. Training employed binary cross-entropy loss optimization via AdamW with matching hyperparameters and cosine annealing scheduling.

1.5. Evaluation of the compared methods

Self-supervised models were trained in a task-agnostic way, and, following this training, the SSL representations were used as inputs to a shallow classifier finetuned on the outcome variable, the visual scoring of the thymus. Shallow classifiers built on each of the SSL strategies were first tuned to find the optimal hyperparameters using the Optuna framework.⁴⁷ The search strategy employs a comprehensive grid-based approach using RandomizedSearchCV with 5-fold stratified cross-validation. The pipeline systematically evaluates combinations of preprocessing scalers (StandardScaler, MinMaxScaler, RobustScaler, or none), feature selection methods (SelectKBest with `f_classif/mutual_info`, SelectPercentile, PCA, TruncatedSVD, VarianceThreshold, and RFE with RandomForest/LogisticRegression), and seven different classifiers (LogisticRegression, RandomForest, GradientBoosting, SVM, KNN, GaussianNB, ExtraTreesClassifier) with their respective hyperparameter grids. Each combination is evaluated using the area under the receiver operating characteristic curve (ROC-AUC) as the primary scoring metric, with the search space dynamically adjusted based on feature dimensionality constraints and memory limitations, ultimately selecting the best-performing pipeline configuration for each feature set. The best-performing configurations for each SSL strategy were compared on a held-out test set. The supervised models were fine-tuned end-to-end and finally evaluated on the test set. We also compared representations extracted from the supervised models, mimicking the same strategy as for the SSL variants.

1.6. Inter-rater analyses and quality control

To ensure the quality of our annotations and to substantiate our methodological choices, we had a team of 2 expert radiologists thoroughly inspect a random sample of 152 CTs, so we could comprehensively assess the validity and reproducibility of our methods. The experts performed 2 tasks: the identification of the spatial coordinates corresponding to the thymus CoM, and the grading of the thymus according to the thymic score scale proposed previously.^{19,20} For the first task, we assessed the agreement between inter-human annotations in a comprehensive analysis, both on axis-level, comparing the raw coordinates summary statistics and respective distances, and on a global level, measuring the pairwise Euclidean distance. Coordinate-level analysis enabled us to detect directional biases or axis-specific variance, while distance-based metrics provided a scalar summary of overall spatial disagreement. Additionally, to control for the quality of our automated localization method, we measured its agreement against one of the raters, on a sample of 1,118 random scans, and visually inspected 1,000 CoM. Both perspectives are necessary to comprehensively assess the reproducibility and validity of the localization procedure. For the second task, the raters graded the thymus on a discrete scale from 0 to 3, as previously explained.^{19,20} However, as our classification model was trained on the binary task to differentiate between fully fatty degenerated thymus, i.e., visual grade 0 (see examples in **Supplementary Figure 7a**), versus at least minimal soft-tissue density thymus, i.e., visual grades 1, 2, and 3 (see examples in **Supplementary Figure 7b**), we assessed agreement based on the binarized score. Individuals who showed non-fat non-thymic-soft-tissue attenuation in the thymic bed, such as fat stranding, edema, or scarring (see examples in **Supplementary Figure 7c**), the raters considered them as visual grade 0, in line with the biological assumption that local changes of the micromilieu impair thymic function.^{15,17,35} Finally, to better understand the anatomy of the thymus, we inspected the distribution of sizes of the thymic beds in our cohort (**Supplementary Figure 6**). This quality assurance study helped us consolidate the validity of the grading system used as the cornerstone of this paper and make important technical decisions, such as defining the correct input size for our deep-learning models.

1.7. Software

The software used in this work can be found in its respective GitHub repository (https://github.com/AIM-Harvard/thymus_health_deeplearning_system.git). Most software was developed in Python3.8, for the model development and deployment, or R statistical software version 4.2.2., for the statistics.

The Python package used comprise:

brotlipy==0.7.0, certifi==2023.5.7, cffi==1.15.0, charset-normalizer==2.0.4, cryptography==39.0.1, fire==0.4.0, idna==3.4, imageio==2.28.1, importlib-resources==5.12.0, lazy_loader==0.2, loguru==0.7.0, mkl-fft==1.3.1, mkl-random==1.2.2, mkl-service==2.4.0, monai==1.1.0, networkx==3.1, nibabel==5.1.0, nptyping==2.5.0, numpy==1.24.3, packaging==23.1, pandas==2.0.1, Pillow==9.4.0, pip==23.0.1, psutil==5.9.5, pycparser==2.21, pynrrd==1.0.0, pyOpenSSL==23.0.0, PySocks==1.7.1, python-dateutil==2.8.2, pytorch-ignite==0.4.12, pytz==2023.3, PyWavelets==1.4.1, PyYAML==6.0, requests==2.29.0, scikit-image==0.20.0, scipy==1.9.1, setuptools==66.0.0, SimpleITK==2.2.1,

six==1.16.0, termcolor==2.3.0, tiff==2023.4.12, torch==1.12.0, torchaudio==0.12.0, torchvision==0.13.0, tqdm==4.65.0, typing_extensions==4.5.0, tzdata==2023.3, urllib3==1.26.15, wheel==0.38.4, zipp==3.15.0, foundation-cancer-image-biomarker==0.0.1a22, imblearn==0.0, ipykernel==6.29.4, optuna-dashboard==0.15.1, pip-chill==1.0.3, wandb==0.17.0.

The R packages used include:

broom (v1.0.5), circlize (v0.4.15), ComplexHeatmap (v2.14.0), ggrepel (v0.9.4), readxl (v1.4.3), tidyr (v1.3.0), forestplot (v3.1.3), abind (v1.4-5), checkmate (v2.3.0), meta (v6.5-0), survminer (v0.4.9), ggpubr (v0.6.0), ggplot2 (v3.4.4), survival (v3.5-7), dplyr (v1.1.4).

Supplementary Methods S2 | Deep Learning Pipeline Evaluation

2.1. Evaluation of deep learning architectures and learning algorithms

First, we ablated the choice of the best self-supervised meta-architecture to be used in the study.

Supplementary Figure 8 displays the AUC scores across different feature extraction methods, via self-supervised pre-training, for the binary classification of the thymic health scores on a held-out test set. The methods are evaluated on this set after hyperparameter tuning via 5-fold cross-validation on the training data. Metrics were generated across the entire dataset by aggregating predictions on each held-out set across 5 folds. The results demonstrate the effectiveness of different self-supervised learning strategies for thymic health prediction. In panel (a), the ablation study reveals that SwAV and MAE achieve significant performance gains over all of the other methods with AUCs of 0.746 and 0.758, respectively. DINOv2 (AUC=0.717, $p<0.001$) and NNCLR (AUC=0.644, $p<0.001$) perform worse than both SwAV and MAE. Panel (b) further validates the superiority of self-supervised learning (SSL) approaches, showing that SSL methods using MAE (ViT backbone) (0.758) and SwAV (ResNet-50 backbone) (0.746) significantly outperform their supervised counterparts with identical backbones (ViT: 0.627, ResNet: 0.554). Moreover, using feature representations and a shallow classifier outperforms end-to-end training of even supervised models (0.664 vs 0.554, $p<0.001$). As SwAV and MAE both show competitive performance, we choose SwAV as the final method for our subsequent analyses due to the simplicity of its convolutional architecture when compared to transformers, along with the advantages posed by CNNs in handling different input sizes and resolutions.

Upon selection of SwAV as the meta-architecture for feature representations with hyperparameters of interest described in (see **Supplementary Table 10**), we further developed our shallow classifier. The hyperparameters for the prediction model developed from these learned representations were selected through 5-fold cross-validation following the exact systematic search strategy for architecture selection ablations detailed in Section 1.5. The prediction model obtained a very good agreement with the manual binary target variable with an AUC-ROC of 0.79 when performance was aggregated across the splits (**Supplementary Figure 8c**). The final model, with the optimal hyperparameters, was retrained on all the data before applying it to further cohorts. We plotted the distribution of prediction probabilities across fine-grained visual thymic health assessments (0, fully fatty degenerated thymus; 1, minimal soft-tissue density thymus; 2, approximately half fatty, half soft-tissue density thymus; 3, predominantly soft-tissue density attenuation thymus), as previously described^{19,20}; and we found that the prediction confidence aligns well with visual grading scales indicating a linear trend (see **Supplementary Figure 8d**). To assess the model's performance for different cancer types, we calculated the AUC-ROC for each cancer type subgroup in the accompanying paper titled "Thymic health is associated with immunotherapy outcomes in patients with cancer".

2.2. Quality control

2.2.1. Inter-rater analyses (human vs human)

In this section, the first analysis we conducted assessed how the two independent expert radiologists visually rated a sample of 152 CT scans, using the proposed^{19,20} thymus score scale. Like in the original paper¹⁹, we tested it using the Cohen Kappa test, which yielded a value of 0.9729, suggesting an almost perfect inter-rater agreement for the binary classification of the thymus score (see **Supplementary Table 11**). This validation was crucial since these annotations were the ones used as labels to fine-tune our pre-trained models for the downstream supervised task of grading the Thymus Health score.

Then, we analyzed the variability between the 2 radiologists for the placement of the thymus CoM.

Supplementary Figure 9 displays the distribution of the pairwise distances for each CoM coordinates between human raters for each axis, and **Supplementary Table 12** summarizes its distances, both axis-specific and Euclidean distance. On average, the 2 raters placed their annotations approximately 8mm apart, with a standard deviation (SD) of roughly 7mm. The longitudinal axis, i.e., head-to-toe, was the direction where raters varied the most. Additionally, **Supplementary Figure 10** shows the quasiperfect agreement of CoM placement between the 2 raters, with intra-class correlation (ICC) coefficients above 0.96 for all axes. **Supplementary Table 13** summarizes their individual statistics for the raw CoM coordinates.

2.2.2. Validation of the automated CoM localizer

To assess the quality of the CoM Localization Model, we conducted a quality control inspection on a random sample of the testing cohort composed of 1,000 CT scans, a subset of the NLST cohort. An expert reader reviewed these 1,000 automatically assessed cases and confirmed excellent accuracy of the automatic CoM detection in 99.8% (998/1,000) of scans (qualitative examples can be found in

Supplementary Figure 11). Additionally, to quantify the precision of the automatically computed CoMs, using the training dataset, we compared 1,118 pairs of manually expert-annotated CoM and their automatic counterparts and found, on aggregate, a deviation from the manual annotation location of -1 mm (-8, 6), on the x-axis, where negative values mean a shift to the left compared to the manual annotation, 0 mm (-6, 6) on the y-axis, where negative values mean an anterior shift, 5 mm (-6, 16) on the z-axis, where negative values mean an upwards shift, and 12.5 mm (4, 20) for the overall euclidean distance between the two annotation approaches (**Supplementary Figure 12**). As shown in the following section, the input stability testing demonstrates that the model performs excellently within this margin of error without compromising performance. This extra step assesses the applicability of the thymus localization model, allowing a fully automated pipeline. This is critical for use cases where manual annotation is either too time-consuming or simply unfeasible, and it facilitates potential clinical implementation for fully automatic opportunistic screening.

2.2.3. Quantification model robustness

As segmentation contours can vary across readers and scan acquisitions, and since the quantification model depends on the input of the location of the CoM of the thymus, it is important to assess the model stability in cases where the input may be imperfect. We extensively tested our system against variations in simulated inter-reader and test-retest scenarios. For the test-retest stability, we used the well-established RIDER⁴⁸ dataset for repeatable radiomics features, composed of 31 pairs of CT scans taken roughly 15 minutes apart and using the same imaging protocol for each patient. We obtained excellent intra-class correlation coefficients (Automatic QC ICC=0.975, Thymic Health ICC=0.970) and excellent Cohen's kappa (Automatic QC=0.86, Thymic Health ICC=0.9) for continuous model predictions and categorical grading, respectively (see **Extended Data Figure 7a-b**). Additionally, inter-reader variations were simulated through 50 perturbation trials where the CoMs were perturbed by sampling from a three-dimensional Gaussian distribution centered at zero with a variance of 16mm. The variance of 16mm was inspired by previous work^{49,50} and confirmed appropriate by the previous results, where we concluded the average Euclidean distance among the sample of 1,118 was 12.5mm. By sampling noise from a normal distribution with a variance of 16mm, i.e., standard deviation (SD) of 4mm, we effectively sample from a range of values between -12mm and 12mm, since 99.7% of the values of a normal distribution lie within 3 SD from the mean. Even though the inter-human variability was smaller for the sample tested in the inter-rater study, i.e., 8mm, we opted for stress-testing the input variation conservatively to cover the fully automated approach. For this analysis, we used the entire FHS, composed of 2,581 scans, and additionally, we randomly sampled 2,500 CT scans from NLST to predict the thymus scores and compared the agreement between model predictions among each of the 50 trials. The concordance index (CI) was used to measure agreement between pairs. We observed excellent performance between scores calculated across pairs of trials for both cohorts, yielding median CI above 0.90 (see **Extended Data Figure 7c-e**).

2.3. AI-generated feature interpretability

To shed light on the underlying dynamics of the deep-learning model and to increase interpretability, we investigated the individual contribution of the 4,096 features to the AI model's output by analyzing its Shapley values.⁵¹ Shapley value analysis is a game theory-inspired method that allows the quantification of individual feature contributions towards the model output. The Shapley value of a feature can be interpreted as the average marginal contribution of that specific feature across all possible permutations of features. In the case of linear models, the Shapley value for a feature can be simplified to the deviation of the feature value from its expected value across the cohort, multiplied by the feature's coefficient. This indicates how much a change in the feature impacts the model's output. **Supplementary Figure 13** shows the distribution of Shapley values across features and datasets in presumed healthy individuals of the FHS and NLST cohorts. To investigate the model's decision-making process, we examined Shapley values and observed that no single feature drives the model's output, but rather the combination of all 4,096 features, observing different features being important for different samples, corroborating the necessity of a high-dimensional space to accurately describe the samples.

Occlusion sensitivity is a method that systematically occludes regions of an input image to understand a neural network's decision-making process. To assess the explainability of the model, we estimate the contributions of different image regions for the final prediction, by selecting the features that contributed the most to the model's prediction for each scan, i.e., larger absolute Shapley values, and computed the saliency maps using an occlusion-sensitivity-based approach. As salient regions are occluded, the deviations in values are larger and indicate the occluded volume's greater significance

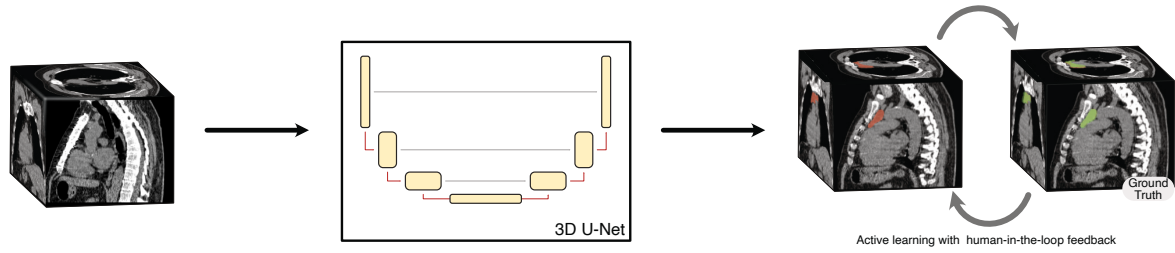
in the network's decision. This process highlights the salient regions of the input image that contribute most to the model's decision. We visualized these regions using saliency maps to make the model's outputs more interpretable (see **Extended Data Figure 8**). As expected, the model focused on the thymic bed and its contours while giving minor attention to the surrounding context, such as the adjacent sternum or lung tissue.

2.4. Pipeline performance

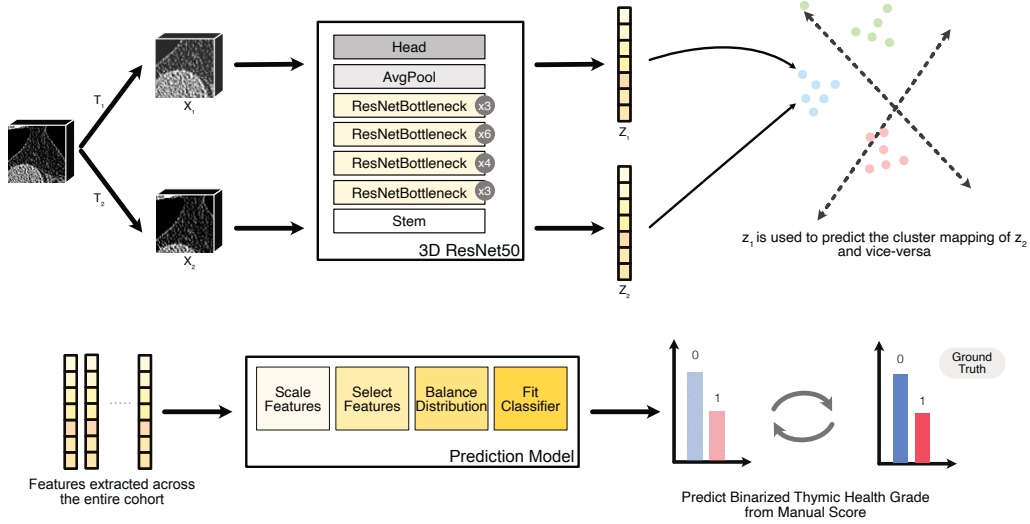
Regarding the performance of the model in our setting, we observed that the end-to-end deep learning system could analyze each scan on average under 2 seconds on a Quadro RTX 8000 48GB GPU system, allowing the automatic assessment of the more than 25,000 CT scans, of average size per scan of 33MB, of the participants of the FHS and NLST within 14 hours.

SUPPLEMENTARY FIGURES

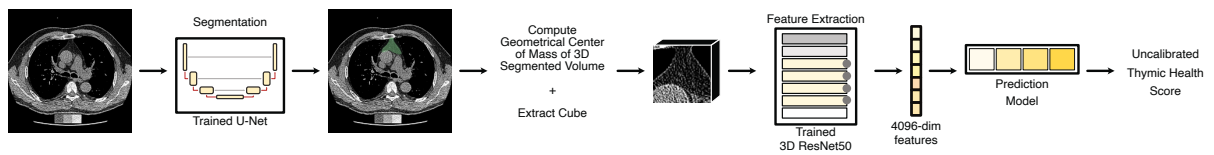
a Training Localization Model



b Training Quantification Model



c End-to-End Deep Learning System for Thymic Health Quantification



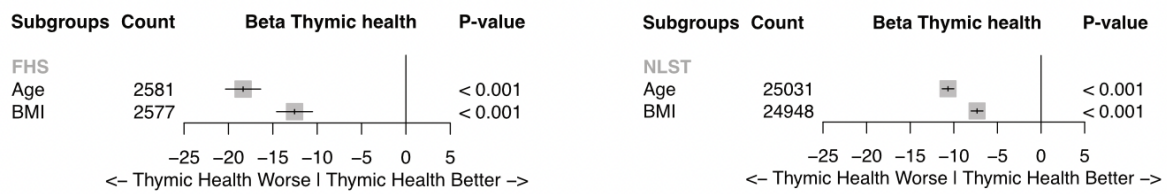
Supplementary Figure 1 | Development of the deep learning system

The deep learning system comprises localization and quantification models, which are trained separately, as shown in **a** and **b**. These models are applied sequentially, resulting in an end-to-end pipeline shown in **c**.

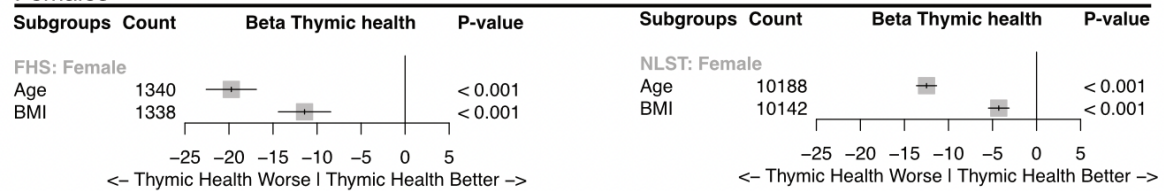
Framingham Heart Study

National Lung Screening Trial

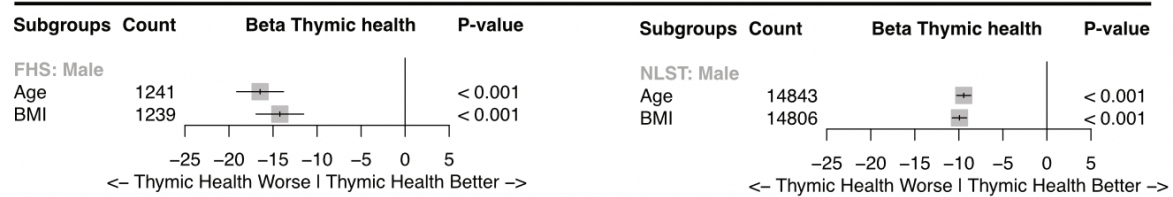
a All individuals



b Females



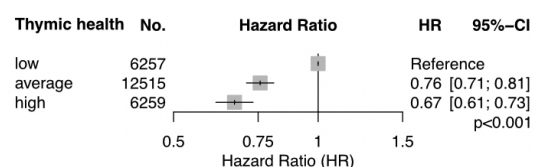
c Males



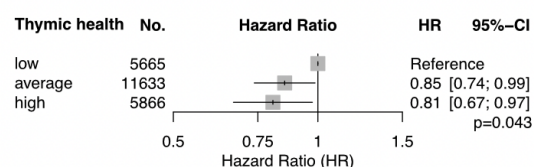
Supplementary Figure 2 | Association of age and body mass index (BMI) with thymic health

Linear regression analysis of age or BMI with thymic health as outcome with **a**, adjustment for sex, and stratified in **b**, female or **c**, male subgroups. Standardization shows the change in mean thymic health for a one standard deviation increase in age or BMI. BMI Body Mass Index, FHS Framingham Heart Study, NLST National Lung Screening Trial.

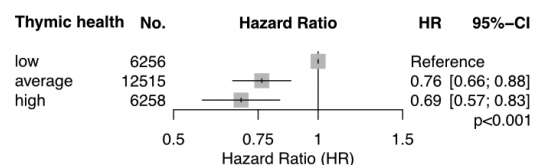
a Overall survival



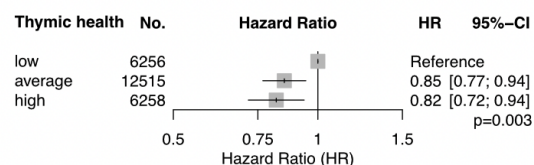
b Lung cancer incidence



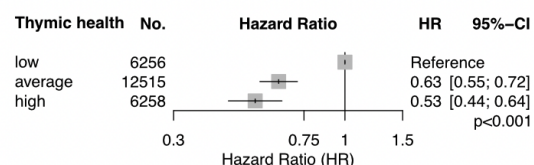
c Lung cancer mortality



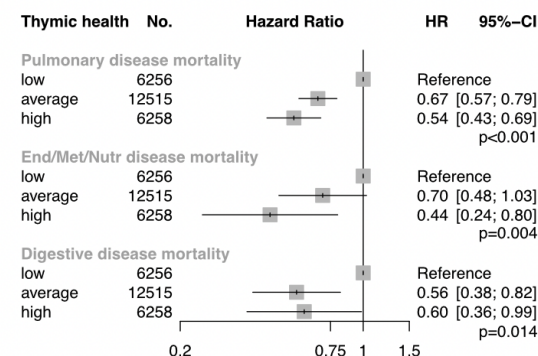
d Malignant neoplasm mortality



e CVD mortality



f Disease specific mortality

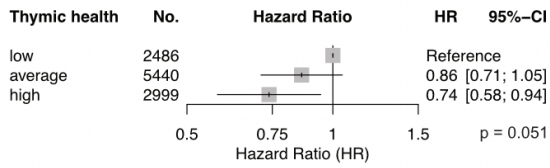


Supplementary Figure 3 | Association of thymic health with health outcomes in NLST using 3-year binned age stratification

Analyses were adjusted for smoking status, continuous pack-years, and stratified by sex and 3-year age bins to account for potential residual confounding by age. Hazard ratios of **a**, overall survival, **b**, lung cancer incidence, **c**, lung cancer mortality, **d**, malignant neoplasm mortality, **e**, cardiovascular-disease specific mortality, and **f**, mortality from the specified disease groups. Follow-up for all analyses was 12 years. Forest plots show type III p-value as test for significance of thymic health, conditional on the other variables present in each multivariate model, respectively. CI Confidence Interval, End/Met/Nutr endocrine, nutritional, and metabolic diseases, HR Hazard Ratio, NLST National Lung Screening Trial.

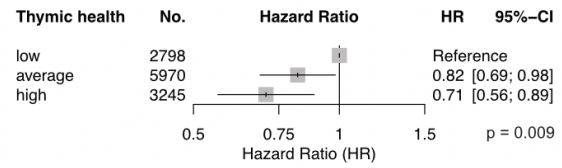
Lung cancer incidence in NLST

a Current smoker

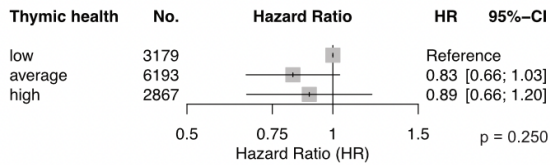


Lung cancer mortality in NLST

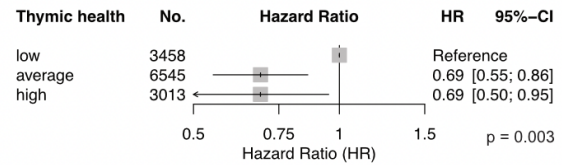
c Current smoker



b Former smoker



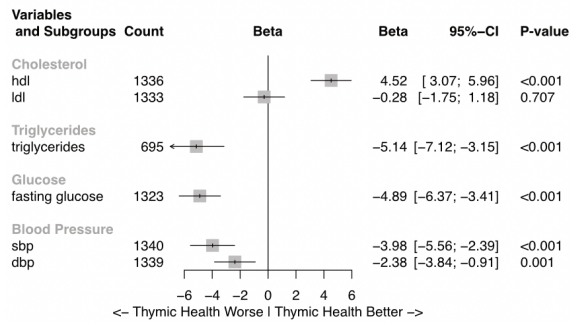
d Former smoker



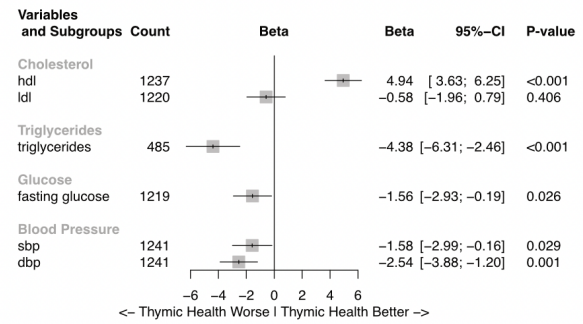
Supplementary Figure 4 | Association of thymic health with long-term risk of lung cancer and lung cancer-specific mortality according to smoking status

Hazard ratio of new lung cancer in NLST participants who were **a**, current or **b**, former smokers, stratified by sex and age, and adjusted for pack-years. Hazard ratio of death from lung cancer in NLST participants who were **c**, current or **d**, former smokers, stratified by sex and age, and adjusted for pack-years. Forest plots show type III p-value as test for significance of thymic health, conditional on the other variables present in each multivariate model, respectively. Former smokers must have quit smoking within 15 years of eligibility determination to have been eligible for the NLST. Lung cancer incidence, current smoker, n = 565; former smoker, n = 408; lung cancer deaths, current smoker, n = 646; former smoker, n = 370. CI Confidence Interval, FHS Framingham Heart Study, HR Hazard Ratio, NLST National Lung Screening Trial.

a FHS: Female

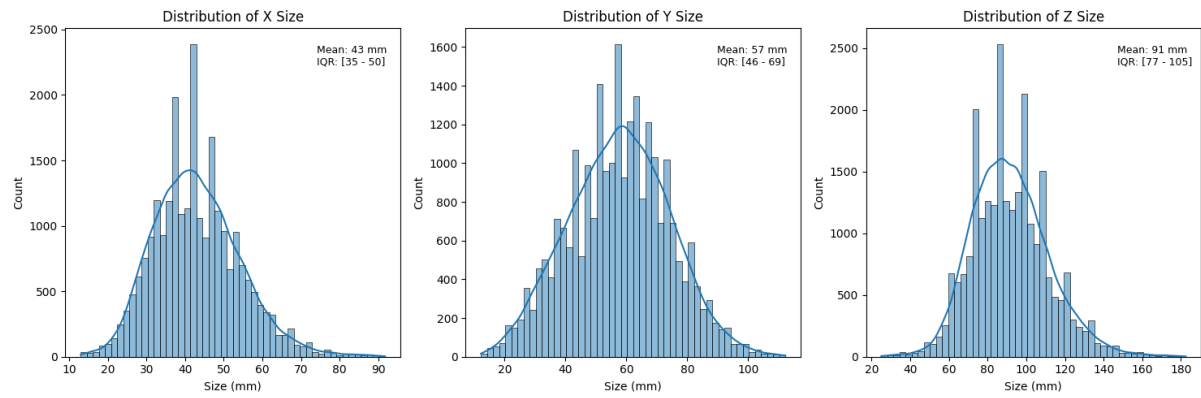


b FHS: Male



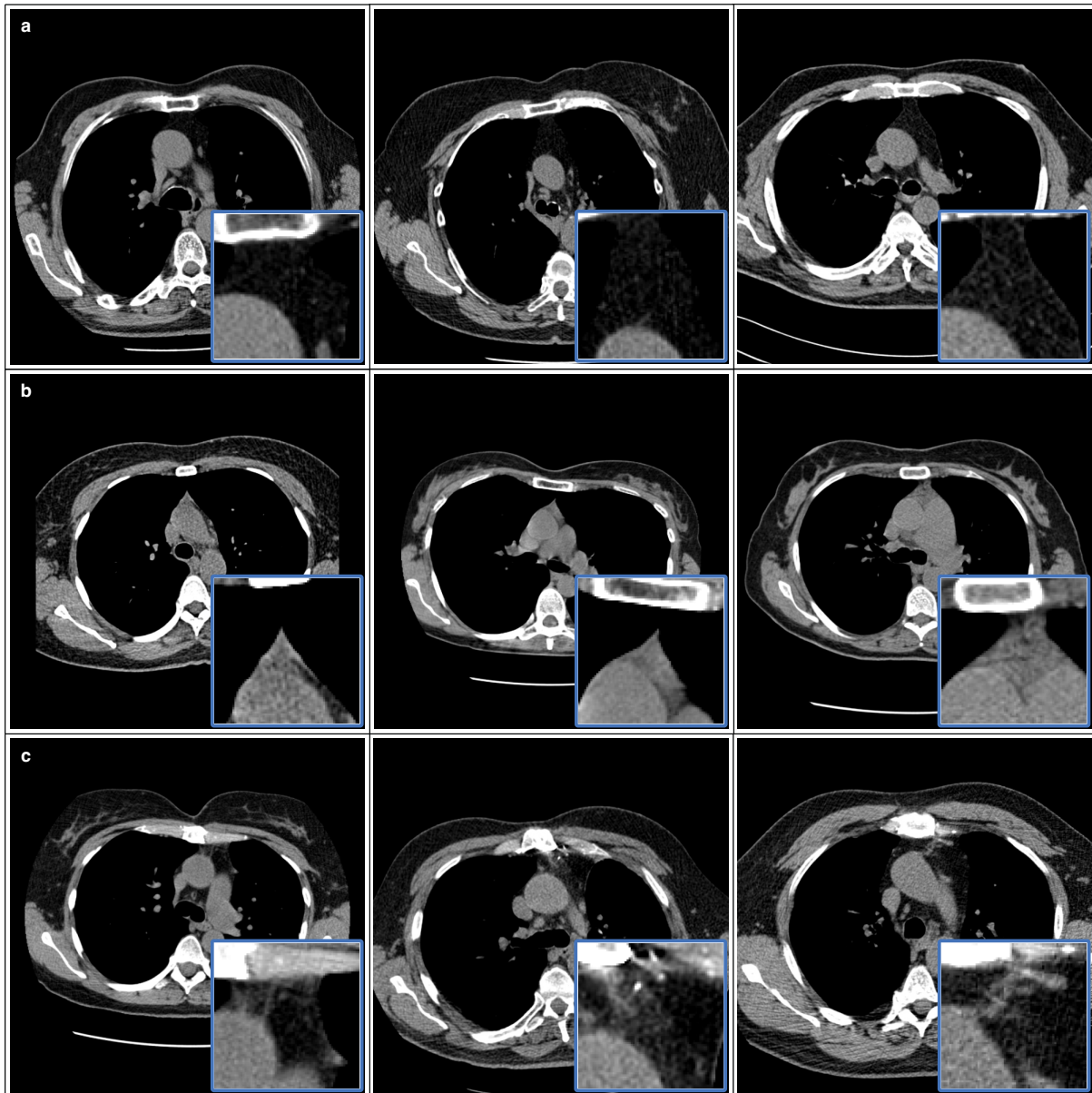
Supplementary Figure 5 | Associations of thymic health with metabolism according to sex

Associations of metabolically relevant variables with thymic health in the Framingham Heart Study stratified by sex into **a**, female, and **b**, male participants, adjusted by age and smoking status. For linear regression analysis, variables were z-score normalized and treated as predictors for thymic health. The beta value indicates the change per standard deviation increase in each predictor. DBP diastolic blood pressure, HDL high-density lipoprotein, LDL low-density lipoprotein, SBP systolic blood pressure.



Supplementary Figure 6 | Distribution of thymic bed sizes along each spatial axis

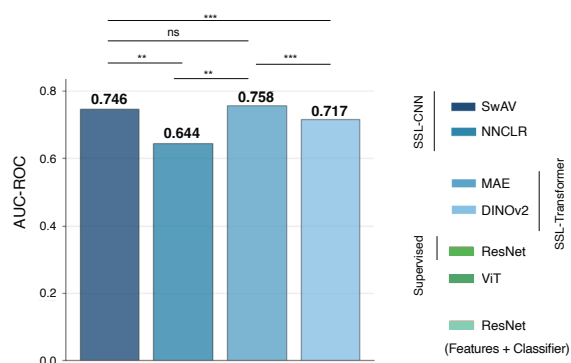
Displays the distribution of the sizes of the thymic bed, per axis, and corresponding probability density function, mean, and interquartile range (IQR) for the NLST population (n=25,031).



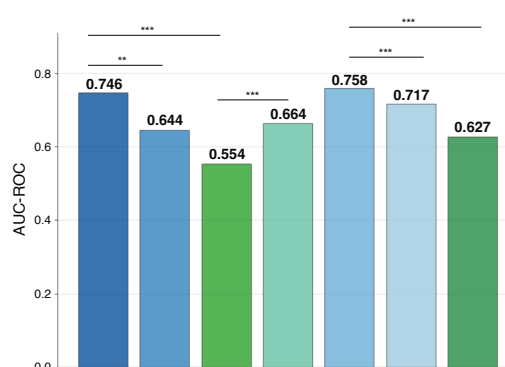
Supplementary Figure 7 | Representative computed tomography (CT) images of manually annotated thymic extreme representations

a, Complete fat attenuation due to fatty degeneration of the thymus. **b**, Non-fat attenuation due to at least residual thymic soft tissue. **c**, Non-fat attenuation due to nonspecific changes such as scarring or inflammation. From left to right, three representative images are shown.

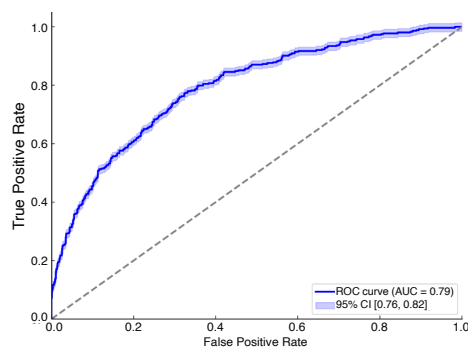
a Ablation of different self-supervised strategies in predicting visual scoring as the downstream task



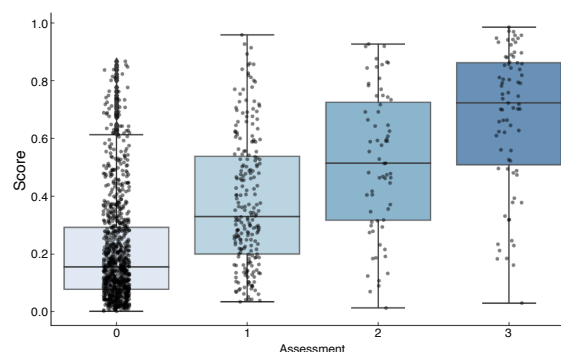
b Comparisons demonstrating the strength of self-supervised methods and feature modeling strategies for the visual scoring downstream task



c ROC curve comparing final thymic health predictions against visual scoring aggregated on five-fold CV split

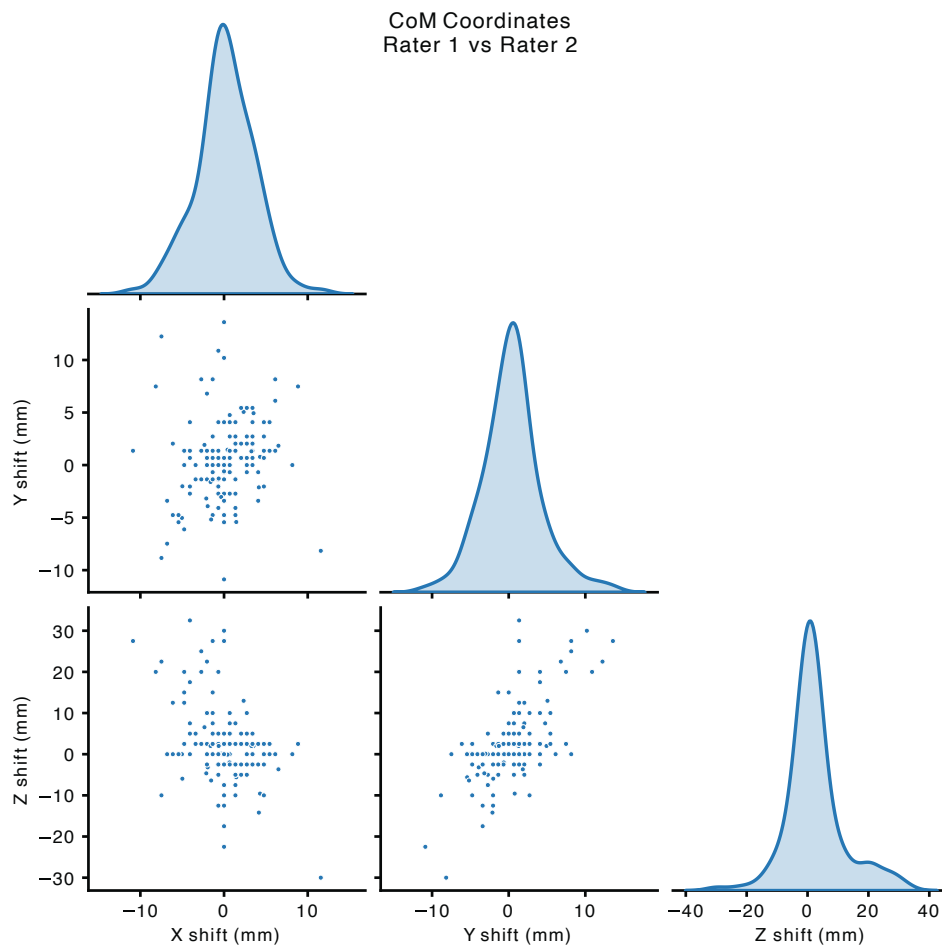


d Distribution of predicted probabilities on the training cohort grouped by fine grained assessment categories (on five-fold CV split)

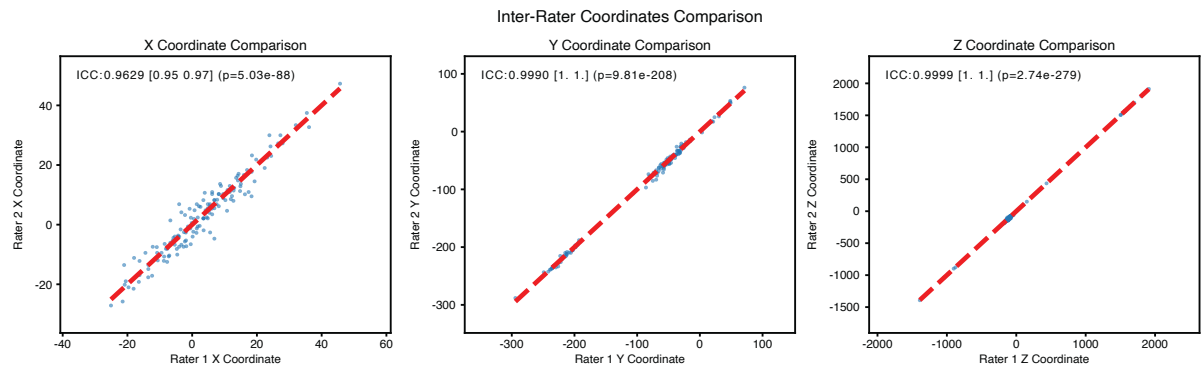


Supplementary Figure 8 | Evaluation of the deep learning system

Components within the thymic health deep learning system are validated independently. First, the selection of self-supervised strategy for the thymus quantification is compared on the visual scoring downstream task, **a**. Area under the ROC-curve is displayed for predictive feature-based models developed using features computed from each of the strategies, on the held-out test set. **b**. Comparisons showing the benefit of self-supervised learning and feature-based modeling vs end-to-end supervised training. The final predictive model was selected using a five-fold CV on the full dataset. In **c**, we compare thymic health predictions against manual grading across these five-fold CV splits. In **d**, we show distribution plots comparing continuous model predictions against finer assessment categories determined during the manual grading process. AUC area under the curve, CV cross-validation, ROC receiver operating characteristic. Group comparisons for panels a and b were performed using a two-sided t-test for two related samples. P values are displayed as significance symbols with *** denoting $p < 0.001$, ** for $p < 0.01$, and ns for non-significant. No adjustment for multiple comparisons was applied.

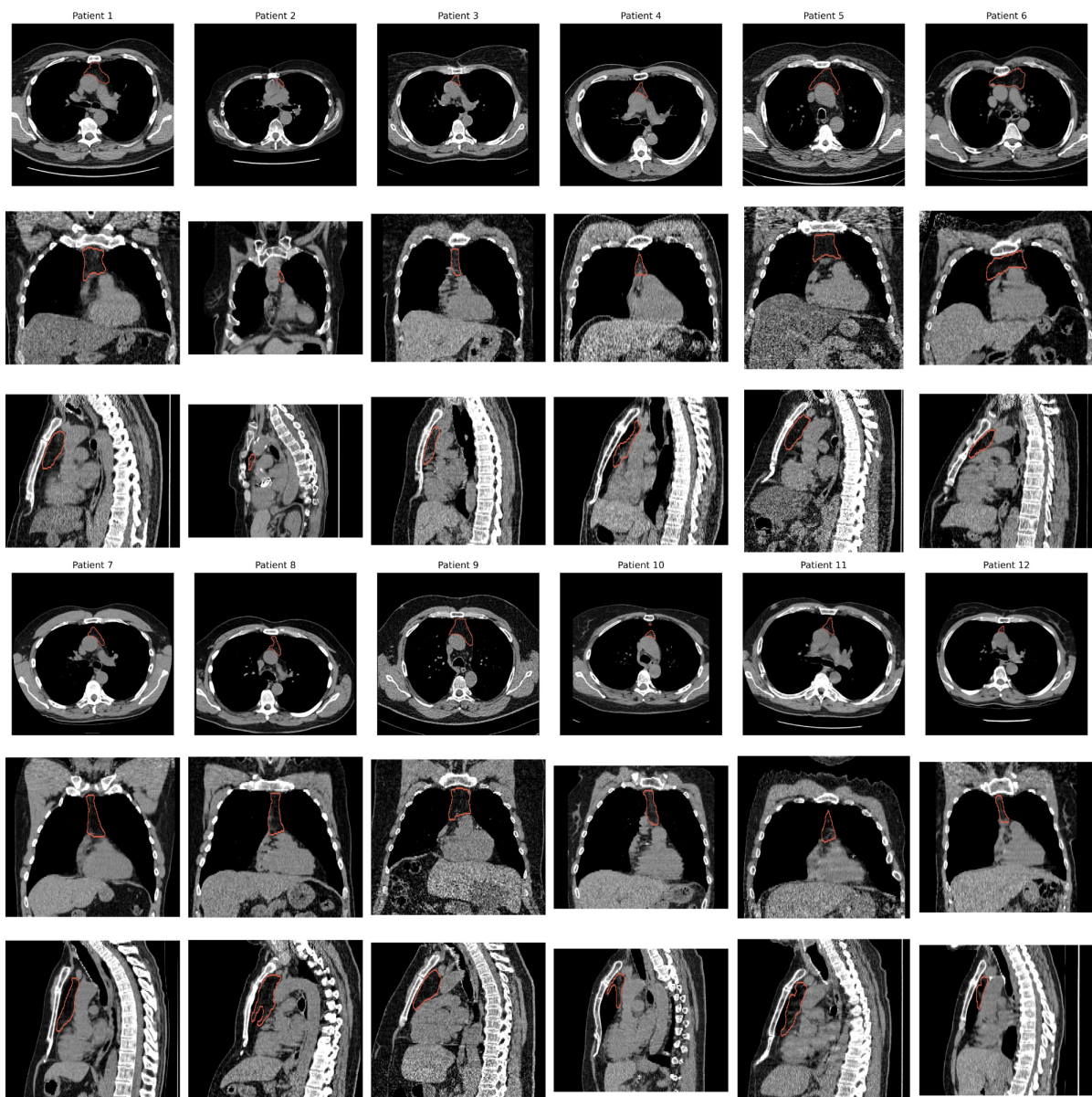


Supplementary Figure 9 | Spatial distributions of pairwise comparisons between CoM annotations per axis
 Displayed are the inter-human distribution of the axis-wise distances between each pair of CoM annotations (n=152) and the respective probability density function.

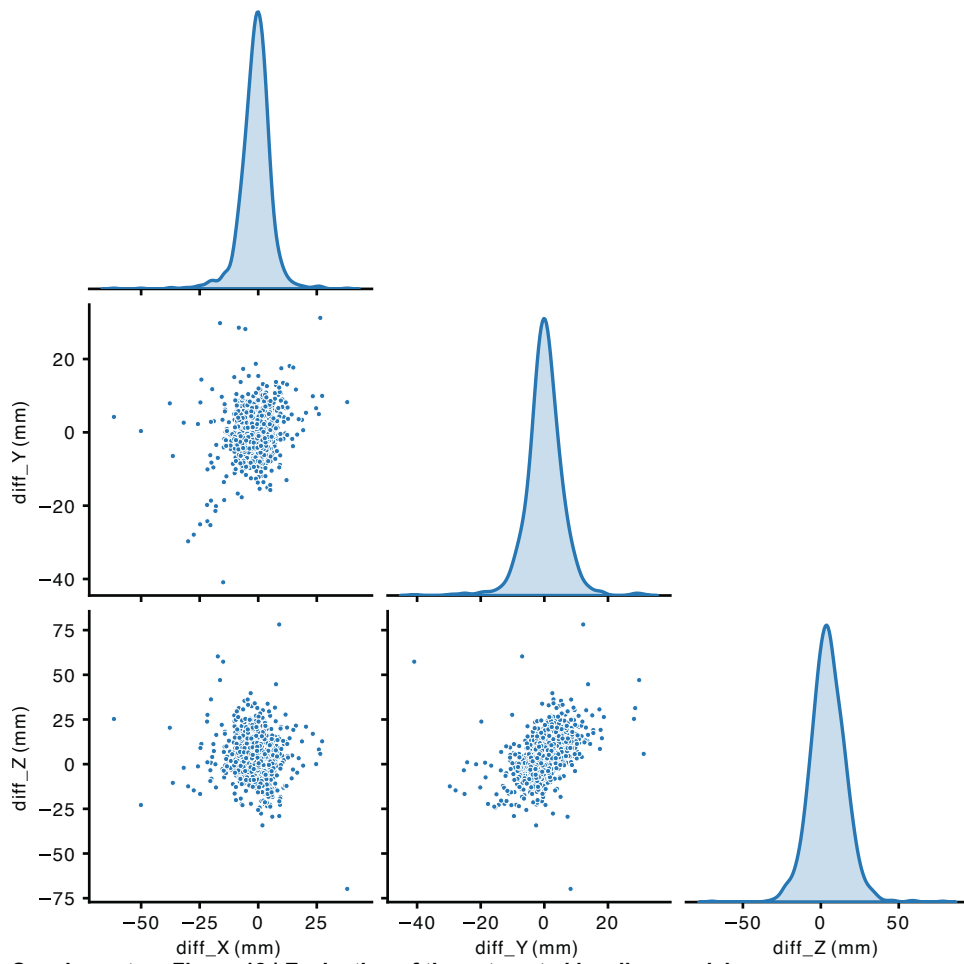


Supplementary Figure 10 | Distribution of thymus sizes along each spatial axis

This figure compares the inter-rater CoM placement ($n=152$) for the 3 axial directions. The red dashed line represents perfect agreement between the 2 raters, and the intra-class correlation coefficient (ICC), the corresponding interquartile range, and p-values are displayed in the upper-left corner of each subplot.



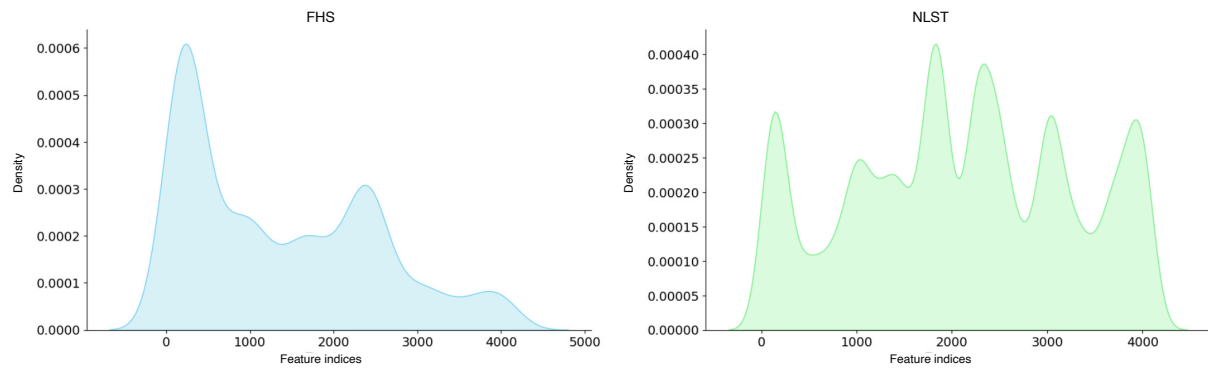
Supplementary Figure 11 | Representative examples of thymic segmentations from the auto-localization model
A total of 12 random patients are shown with axial, coronal, and sagittal reconstruction of the chest computed tomography scan. Visualization of the fully automatic thymic bed segmentation is shown as a red outline.



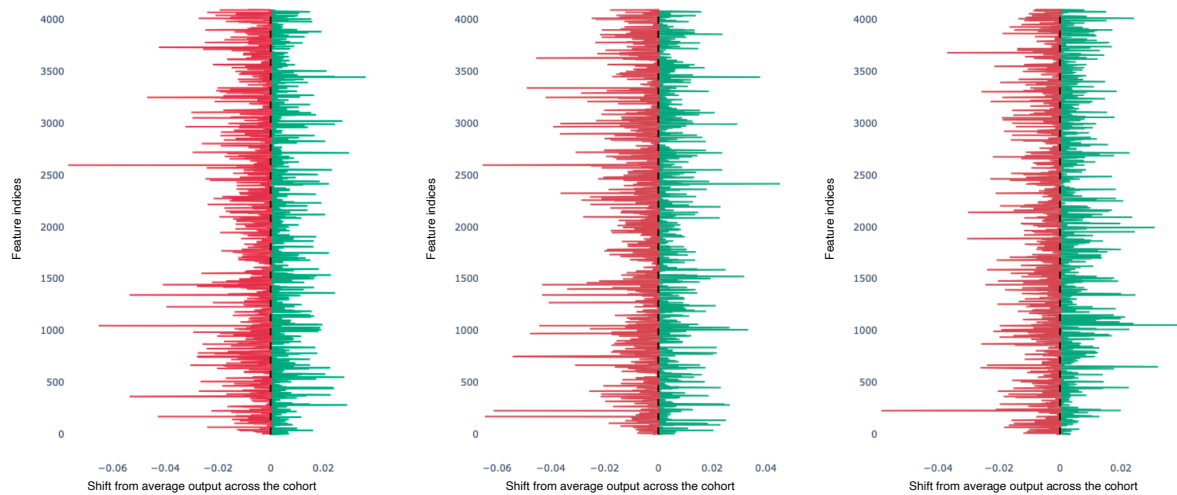
Supplementary Figure 12 | Evaluation of the automated localizer model

The localization component of the system is validated by analyzing differences between the manually annotated CoM seed points and those computed using the predicted segmentations.

a Distribution of the features across the cohort, indicating the frequency of the feature selected as the most important feature using Shapley values.

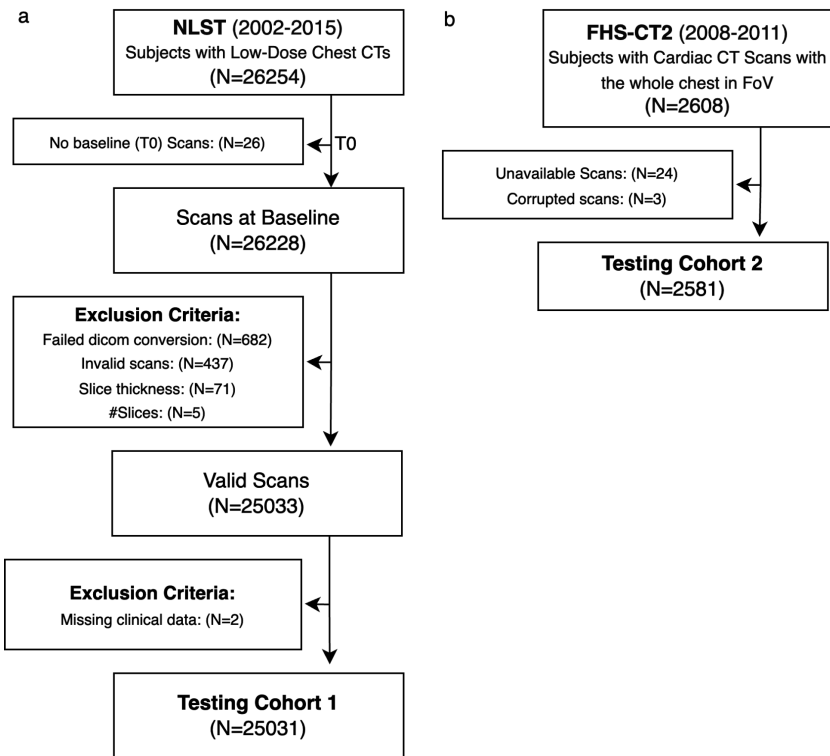


b Shapley value distributions for each of the features corresponding to the representative saliency maps in Extended Data Figure 8.



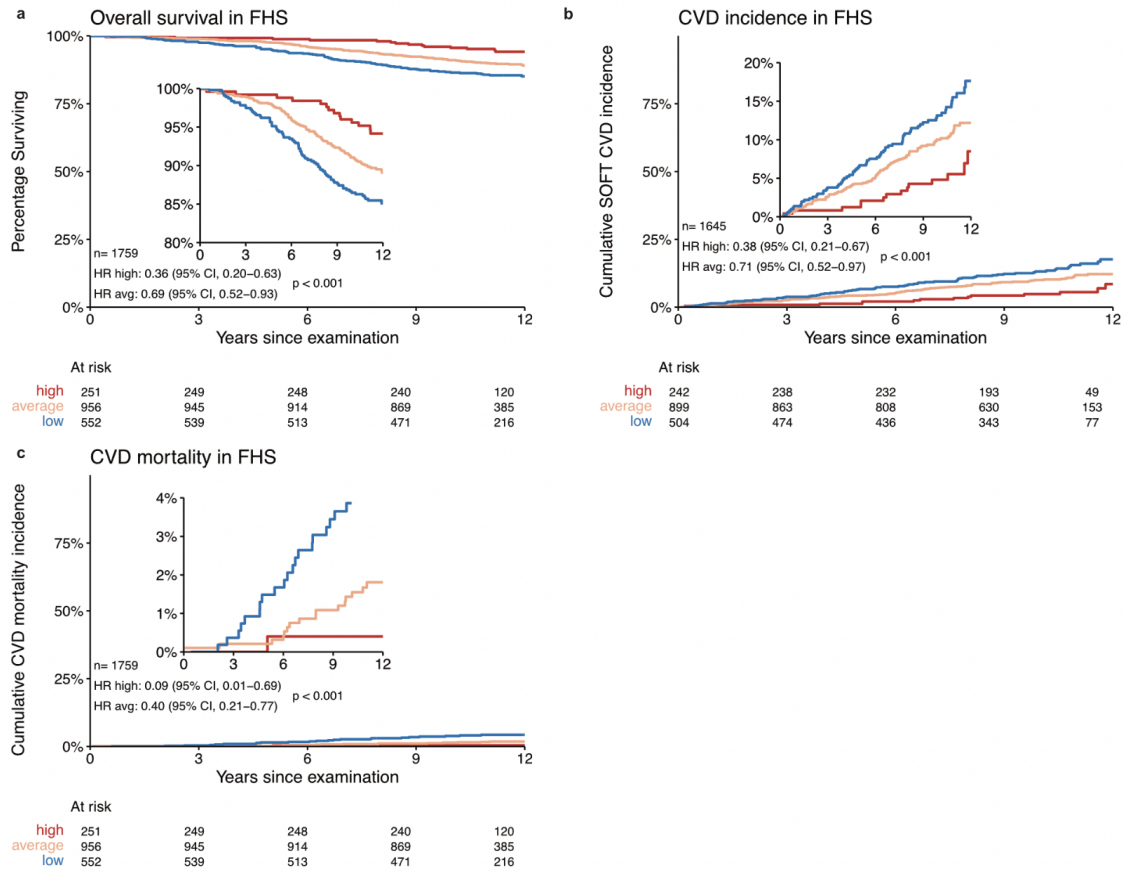
Supplementary Figure 13 | Feature explainability of thymic health on FHS and NLST cohorts

a, The most important features for predicting each thymus grade, as determined by Shapley value analysis of the 4,096 features used by the self-supervised model. **b**, The Shapley value distribution corresponding to the 3 representative CT scans displayed in Extended Data Figure 8. The distribution indicates how features shift the prediction to low or high thymic health from the cohort's average.



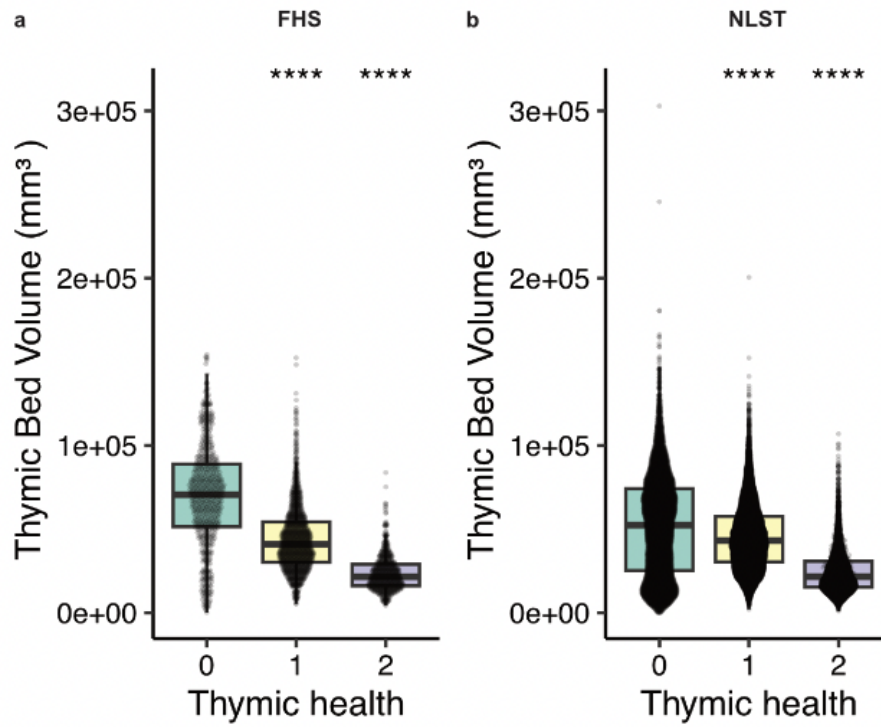
Supplementary Figure 14 | Study population inclusion

Consort diagrams of the testing datasets. **a**, Framingham Heart Study (FHS-CT2), composed of cardiac CTs acquired between 2008-2011, and **b**, National Lung Study Trial (NLST), a low-dose CT collection gathered between 2002-2015.



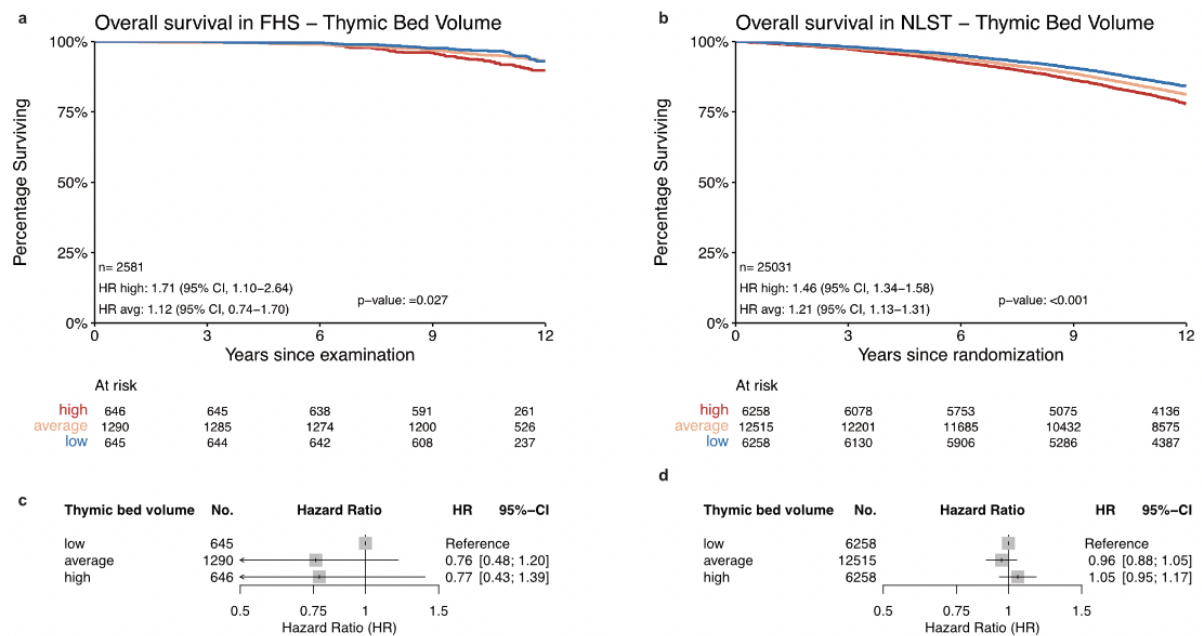
Supplementary Figure 15 | Sensitivity analysis of thymic health according to independent manual thymus grading

From the Framingham Heart Study, we analyzed individuals previously scored as having fully fatty degenerated thymus (category 0) by Araki et al.¹⁹ (2016, $n = 1,759$). Thymic health stratified these individuals into low, average, and high thymic health with $n = 552$, $n=956$, and $n=251$, respectively. **a**, Overall survival, **b**, participants with cardiovascular disease (CVD) events, such as myocardial infarction, congestive heart failure, or cerebral embolism, and **c**, participants who did not die from CVD. Follow-up for all analyses was 12 years. Type III p -value as test for significance of thymic health is shown. CI Confidence Interval, FHS Framingham Heart Study, HR Hazard Ratio.



Supplementary Figure 16 | Association of thymic health with thymic bed volume

We extracted the volume segmentation metrics from our fully automatic thymic bed segmentation model to investigate associations with thymic health. FHS Framingham Heart Study, NLST National Lung Screening Trial. Group comparisons were performed using the Wilcoxon rank-sum test with two-sided P values. P values are displayed as significance symbols with **** denoting $p < 0.0001$. No adjustment for multiple comparisons was applied.



Supplementary Figure 17 | Association of thymic bed volume with long-term mortality

Shows weak associations in adjusted analyses which do not hold after adjustment adding evidence that thymic health captures compositional characteristics of the thymus extending beyond thymic volume. **a**, Survival of participants from the Framingham Heart Study (FHS) and **b**, the National Lung Screening Trial (NLST). **c**, Hazard ratio of death from any cause for FHS participants adjusted for smoking status and stratified by sex and age. **d**, Hazard ratio of death from any cause for NLST participants adjusted for pack-years and smoking status and stratified by sex and age. Follow-up for all analyses was 12 years. CI Confidence Interval, FHS Framingham Heart Study, HR Hazard Ratio, NLST National Lung Screening Trial.

SUPPLEMENTARY TABLES

Supplementary Table 1 | Distribution of diagnoses for malignant neoplasm-specific death

Description	ICD Code	Count	%
Malignant neoplasm of bronchus and lung	C34	1,075	52
Malignant neoplasm of pancreas	C25	153	7
Malignant neoplasm of oesophagus	C15	67	3
Malignant neoplasm, without specification of site	C80	64	3
Malignant neoplasm of colon	C18	62	3
Malignant neoplasm of prostate	C61	53	3
Malignant neoplasm of bladder	C67	52	3
Malignant neoplasm of liver and intrahepatic bile ducts	C22	50	2
Myeloid leukaemia	C92	47	2
Malignant neoplasm of brain	C71	44	2
Malignant neoplasm of breast	C50	40	2
Malignant neoplasm of kidney, except renal pelvis	C64	38	2
Multiple myeloma and malignant plasma cell neoplasms	C90	31	2
Other and unspecified types of non-Hodgkin lymphoma	C85	30	1
Malignant neoplasm of ovary	C56	20	1
Malignant neoplasm of stomach	C16	20	1
Malignant neoplasm of larynx	C32	17	1
Malignant melanoma of skin	C43	14	1
Non-follicular lymphoma	C83	11	1
Leukaemia of unspecified cell type	C95	11	1

Footnotes: Detailed causes of death from the ICD-10 codes included in the malignant neoplasm-specific survival analysis of participants of the National Lung Screening Trial are displayed, excluding those that constituted under one percent of all diagnoses. ICD-10 International Statistical Classification of Diseases 10th revision.

Supplementary Table 2 | Distribution of diagnoses for cardiovascular-specific death

Description	ICD Code	Count	%
Chronic ischaemic heart disease	I25	381	33
Acute myocardial infarction	I21	213	18
Stroke, not specified as haemorrhage or infarction	I64	65	6
Heart failure	I50	52	4
Intracerebral haemorrhage	I61	40	3
Hypertensive heart disease	I11	40	3
Aortic aneurysm and dissection	I71	39	3
Cardiac arrest	I46	37	3
Cardiomyopathy	I42	30	3
Other nontraumatic intracranial haemorrhage	I62	25	2
Essential (primary) hypertension	I10	22	2
Complications and ill-defined descriptions of heart disease	I51	22	2
Nonrheumatic aortic valve disorders	I35	18	2
Atrial fibrillation and flutter	I48	17	1
Cerebral infarction	I63	17	1
Pulmonary embolism	I26	15	1
Other pulmonary heart diseases	I27	15	1
Sequelae of cerebrovascular disease	I69	14	1
Other peripheral vascular diseases	I73	12	1
Other cardiac arrhythmias	I49	11	1
Atherosclerosis	I70	10	1
Subarachnoid haemorrhage	I60	9	1
Other cerebrovascular diseases	I67	8	1
Other acute ischaemic heart diseases	I24	8	1
Hypertensive renal disease	I12	6	1

Footnotes: Detailed causes of death from the ICD-10 codes included in the cardiovascular-specific survival analysis of participants of the National Lung Screening Trial are displayed, excluding those that constitute under one percent of all diagnoses. ICD-10 International Statistical Classification of Diseases 10th revision.

Supplementary Table 3 | Distribution of diagnoses for pulmonary disease-specific death

Description	ICD Code	Count	%
Other chronic obstructive pulmonary disease	J44	517	71
Other interstitial pulmonary diseases	J84	58	8
Pneumonia, organism unspecified	J18	42	6
Emphysema	J43	37	5
Pneumonitis due to solids and liquids	J69	25	3
Respiratory failure, not elsewhere classified	J96	14	2
Other respiratory disorders	J98	9	1
Bacterial pneumonia, not elsewhere classified	J15	6	1
Adult respiratory distress syndrome	J80	4	1

Footnotes: Detailed causes of death from the ICD-10 codes included in the pulmonary disease-specific survival analysis of participants of the National Lung Screening Trial are displayed, excluding those that constitute under one percent of all diagnoses. ICD-10 International Statistical Classification of Diseases 10th revision.

Supplementary Table 4 | Distribution of diagnoses for endocrine, nutritional, and metabolic disease-specific death

Description	ICD Code	Count	%
Unspecified diabetes mellitus	E14	70	52
Disorders of lipoprotein metabolism and other lipidaemias	E78	18	13
Type 2 diabetes mellitus	E11	14	10
Obesity	E66	8	6
Other disorders of fluid, electrolyte and acid-base balance	E87	5	4
Type 1 diabetes mellitus	E10	4	3
Amyloidosis	E85	4	3
Other metabolic disorders	E88	3	2
Other disorders of pancreatic internal secretion	E16	2	1
Unspecified protein-energy malnutrition	E46	2	1
Volume depletion	E86	1	1
Disorders of mineral metabolism	E83	1	1
Unspecified severe protein-energy malnutrition	E43	1	1
Retarded development following protein-energy malnutrition	E45	1	1
Hypofunction and other disorders of pituitary gland	E23	1	1

Footnotes: Detailed causes of death from the ICD-10 codes included in the endocrine, nutritional, and metabolic disease-specific survival analysis of participants of the National Lung Screening Trial are displayed. ICD-10 International Statistical Classification of Diseases 10th revision.

Supplementary Table 5 | Distribution of diagnoses for digestive disease-specific death

Description	ICD Code	Count	%
Alcoholic liver disease	K70	26	19
Fibrosis and cirrhosis of liver	K74	23	17
Paralytic ileus and intestinal obstruction without hernia	K56	11	8
Diverticular disease of intestine	K57	8	6
Cholelithiasis	K80	7	5
Other diseases of digestive system	K92	7	5
Vascular disorders of intestine	K55	7	5
Hepatic failure, not elsewhere classified	K72	6	4
Other diseases of intestine	K63	5	4
Other diseases of liver	K76	4	3
Other diseases of oesophagus	K22	3	2
Acute pancreatitis	K85	3	2
Peptic ulcer, site unspecified	K27	3	2
Gastric ulcer	K25	3	2
Ulcerative colitis	K51	3	2
Other inflammatory liver diseases	K75	2	1
Peritonitis	K65	2	1
Other diseases of pancreas	K86	2	1
Other diseases of biliary tract	K83	1	1
Other diseases of gallbladder	K82	1	1
Cholecystitis	K81	1	1
Duodenal ulcer	K26	1	1
Other disorders of teeth and supporting structures	K08	1	1
Other diseases of stomach and duodenum	K31	1	1
Unspecified abdominal hernia	K46	1	1
Gastritis and duodenitis	K29	1	1
Intestinal malabsorption	K90	1	1

Footnotes: Detailed causes of death from the ICD-10 codes included in the digestive disease-specific survival analysis of participants of the National Lung Screening Trial are displayed. ICD-10 International Statistical Classification of Diseases 10th revision.

Supplementary Table 6 | Fried frailty phenotype analysis in Framingham Heart Study participants

	Fried frailty phenotype subcohort
Sample Size, n	1,183
Female, n (%)	697 (58.9%)
Age, mean (range), (sd)	64.8 (40, 87) sd=8.3
Time difference between CT2 and Exam 8 (Fried frailty), years, mean (range; sd)	3.8 (1.5 to 5.9; 0.8)
Current smoker, n (%)	68 (5.7%)

Footnotes: Participants from exam 8 who had Fried frailty phenotyping performed were evaluated.

Supplementary Table 7 | OLINK Inflammation panel analysis in Framingham Heart Study participants

OLINK proteomics subcohort	
Sample Size, n	317
Female, n (%)	183 (57.7%)
Age, mean (range), (sd)	58.3 (42, 80) sd=8.0
Time difference between CT2 and Exam 7 (protein), years, mean (range; sd)	10.4 (8.1 to 12.2; 0.8) Exam 7
CMV positive, n (%)	144 (45.4%)
Current smoker, n (%)	30 (9.5%)

Footnotes: Participants from exam 7 who had Olink panel analyses performed on blood samples were evaluated.

Supplementary Table 8 | Chronic inflammation analysis in Framingham Heart Study participants

Chronic inflammation subcohort	
Sample Size, n	1,156
Female, n (%)	679 (58.7%)
Age, mean (range), (sd)	58.2 (33, 80) sd=8.3
Time difference between CT2 and Exam 7 (protein), years, mean (range; sd)	10.4 (7.8 to 12.4; 0.8) Measured exams 7-9
Current smoker, n (%)	93 (8.0%)

Footnotes: Chronic systemic inflammation was defined as consistently high CRP levels of greater than or equal to 3 mg/L in at least two examination cycles comprising exams seven to nine.

Supplementary Table 9 | Data collections and respective link

Collection Name	Link
acrin_nslc_fdg_pet	https://doi.org/10.7937/tcia.2019.30ilqfcl
anti_pd_1_lung	https://doi.org/10.7937/tcia.2019.zijwb9ip
cmb_1ca	https://doi.org/10.7937/3cx3-s132
covid_19_ar	https://www.doi.org/10.7937/tcia.2020.py71-5978
covid_19_ny_sbu	https://doi.org/10.7937/TCIA.BBAG-2923
cptac_lsc	https://doi.org/10.7937/K9/TCIA.2018.6EMUB5L2
cptac_luad	https://doi.org/10.7937/k9/tcia.2018.pat12tbs
ct_vs_pet_ventilation_imaging	https://doi.org/10.7937/3PPX-7S22
lctsc	https://doi.org/10.7937/K9/TCIA.2017.3R3FVZ08
lisc_idri	https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX
lung_fused_ct_pathology	https://doi.org/10.7937/k9/tcia.2018.smt36lgn
lung_pet_ct_dx	https://doi.org/10.7937/TCIA.2020.NNC2-0461
lungct_diagnosis	https://doi.org/10.7937/K9/TCIA.2015.A6V7JIWX
midrc_ricord_1a	https://doi.org/10.7937/VTW4-X588
midrc_ricord_1b	https://doi.org/10.7937/31V8-4A40
nslc_radiogenomics	https://doi.org/10.7937/K9/TCIA.2017.7hs46erv
nslc_radiomics	https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI
nslc_radiomics_genomics	https://doi.org/10.7937/K9/TCIA.2015.L4FRET6Z
nslc_radiomics_interobserver1	https://doi.org/10.7937/tcia.2019.cwvlpd26
qin_lung_ct	https://doi.org/10.7937/K9/TCIA.2015.NPGZYBZ
rider_lung_ct	https://doi.org/10.7937/k9/tcia.2015.u1x8a5nr
rider_lung_pet_ct	https://doi.org/10.7937/k9/tcia.2015.ofip7tvm
spie_aapm_lung_ct_challenge	https://doi.org/10.7937/K9/TCIA.2015.UZLSU3FL
tcga_luad	https://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5
tcga_lusc	https://doi.org/10.7937/K9/TCIA.2016.TYGKKFMQ

Supplementary Table 10 | Details of hyperparameters used during the self-supervised pre-training process

Hyperparameters	Values
Batch-Size	512 (effective)
Optimizer	SGD
Learning Rate	0.3
Weight Decay	1e-6
Momentum	0.9
Temperature	0.1
Sinkhorn epsilon	0.03
Number of prototypes	300
Frozen steps at start	50
Queue	Disabled
Epochs	800

Supplementary Table 11 | Confusion matrix for the inter-observer agreement between two expert radiologists

		Observer 2				
		0	1	2	3	Total
Observer 1	0	88	2	0	0	90
	1	0	36	4	0	40
	2	0	0	13	1	14
	3	0	0	3	5	8
	Total	88	38	20	6	152

Supplementary Table 12 | Summary statistics of inter-rater CoM distances

	Mean (mm)	Stand. Dev.	Min	25%	50%	75%	Max
X	2.6	2.3	0	0.7	2.0	4.0	11.6
Y	2.8	2.6	0	0.7	2.0	4.0	13.6
Z	6.0	7.4	0	2.0	2.5	7.5	32.5
Euclidean Distance	8.1	7.2	0.6	3.5	5.7	10.1	33.2

Supplementary Table 13 | Summary statistics of raw CoM coordinates

		Mean (mm)	Stand. Dev.	Min	25%	50%	75%	Max
X	Rater1	2.949	12.590	-25.080	-5.356	2.438	10.813	45.680
	Rater2	2.762	12.908	-27.120	-6.740	2.430	10.260	47.289
Y	Rater1	-93.153	87.480	-294.218	-198.483	-55.720	-36.740	71.125
	Rater2	-93.602	87.469	-288.098	-198.823	-56.132	-38.440	76.156
Z	Rater1	54.442	628.202	-1390.100	-142.125	-112.500	-80.000	1910.500
	Rater2	52.009	628.846	-1392.600	-145.000	-115.250	-80.000	1913.000

SUPPLEMENTARY REFERENCES |

15. Nunes-Alves, C., Nobrega, C., Behar, S. M. & Correia-Neves, M. Tolerance has its limits: how the thymus copes with infection. *Trends Immunol.* **34**, 502–510 (2013).
17. Wang, W., Thomas, R., Sizova, O. & Su, D. M. Thymic Function Associated With Cancer Development, Relapse, and Antitumor Immunity – A Mini-Review. *Front. Immunol.* **11**, 1–9 (2020).
19. Araki, T. et al. Normal thymus in adults: appearance on CT and associations with age, sex, BMI and smoking. *Eur. Radiol.* **26**, 15–24 (2016).
20. Ackman, J. B. et al. Sex difference in normal thymic appearance in adults 20–30 years of age. *Radiology* **268**, 245–253 (2013).
35. Duah, M. et al. Thymus degeneration and regeneration. *Front. Immunol.* **12**, 706244 (2021).
41. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).
42. Diaz-Pinto, A. et al. MONAI Label: A framework for AI-assisted Interactive Labeling of 3D Medical Images. *arXiv [cs.HC]* (2022).
43. Oquab, M. et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv [cs.CV]* (2023).
44. He, K. et al. Masked Autoencoders Are Scalable Vision Learners. *arXiv [cs.CV]* (2021).
45. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P. & Zisserman, A. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. *arXiv [cs.CV]* (2021).
46. Caron, M. et al. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv [cs.CV]* (2020).
47. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (ACM, New York, NY, USA, 2019)*. doi:10.1145/3292500.3330701.
48. Zhao, B. et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* **252**, 263–272 (2009).
49. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* **6**, 354–367 (2024).
50. Hosny, A. et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
51. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017).