scientific reports

Check for updates

OPEN

Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: a multimodal explainable artificial intelligence approach

Juan M. Zambrano Chaves[1], Andrew L. Wentland[2], Arjun D. Desai[3,4], Imon Banerjee[5], Gurkiran Kaur[5], Ramon Correa[5], Robert D. Boutin[3], David J. Maron[6,7], Fatima Rodriguez[6], Alexander T. Sandhu[6], Daniel Rubin[1,3], Akshay S. Chaudhari[1,3,8,9] & Bhavik N. Patel[5,9]✉

Current risk scores using clinical risk factors for predicting ischemic heart disease (IHD) events—the leading cause of global mortality—have known limitations and may be improved by imaging biomarkers. While body composition (BC) imaging biomarkers derived from abdominopelvic computed tomography (CT) correlate with IHD risk, they are impractical to measure manually. Here, in a retrospective cohort of 8139 contrast-enhanced abdominopelvic CT examinations undergoing up to 5 years of follow-up, we developed multimodal opportunistic risk assessment models for IHD by automatically extracting BC features from abdominal CT images and integrating these with features from each patient's electronic medical record (EMR). Our predictive methods match and, in some cases, outperform clinical risk scores currently used in IHD risk assessment. We provide clinical interpretability of our model using a new method of determining tissue-level contributions from CT along with weightings of EMR features contributing to IHD risk. We conclude that such a multimodal approach, which automatically integrates BC biomarkers and EMR data, can enhance IHD risk assessment and aid primary prevention efforts for IHD. To further promote research, we release the Opportunistic L3 Ischemic heart disease (OL3I) dataset, the first public multimodal dataset for opportunistic CT prediction of IHD.

Ischemic heart disease (IHD) is the leading cause of global mortality and among the top causes of morbidity. In 2019, it was responsible for over 9 million deaths worldwide and the loss of more than 180 million disability-adjusted life years (http://ghdx.healthdata.org/gbd-results-tool). Preventive treatments including lifestyle modifications and pharmacologic interventions (e.g., cholesterol-lowering medications) can be guided by risk assessment. The Framingham coronary heart disease risk score (FRS) and the Pooled Cohort Equations (PCE) are commonly utilized risk estimation methods for IHD and atherosclerotic cardiovascular disease, respectively[1,2]. The FRS uses demographic risk factors and cholesterol values to predict 10-year IHD risk in individuals aged

[1]Department of Biomedical Data Science, Stanford University, 1265 Welch Road, MSOB West Wing, Third Floor, Stanford, CA 94305, USA. [2]Department of Radiology, University of Wisconsin-Madison, 600 Highland Ave, Madison, WI 53792, USA. [3]Department of Radiology, School of Medicine, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA. [4]Department of Electrical Engineering, Stanford University, 350 Jane Stanford Way, Stanford, CA 94305, USA. [5]Department of Radiology, Mayo Clinic, 13400 East Shea Blvd, Scottsdale, AZ 85259, USA. [6]Division of Cardiovascular Medicine, Department of Medicine, School of Medicine, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA. [7]Department of Medicine, Stanford Prevention Research Center, School of Medicine, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA. [8]Cardiovascular Institute, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA. [9]These authors contributed equally: Akshay S. Chaudhari and Bhavik N. Patel. ✉email: patel.bhavik@mayo.edu

30–74 years old without known IHD at baseline examination. The PCE were developed to model the 10-year risk of major atherosclerotic cardiovascular disease events, including fatal and nonfatal IHD as well as fatal and nonfatal stroke. These risk scores have been used as a standard for IHD risk assessment in current clinical practice guidelines and policy recommendations, including the most recent American College of Cardiology/American Heart Association guideline on primary prevention of cardiovascular disease[3].

Validation of both risk scores has shown varying performance depending on the subpopulation analyzed. Performance is typically reported as a c-statistic value, which corresponds to the proportion of case–control pairs in which a higher risk is assigned to the case (a measure of discrimination). Previously reported c-statistic values for the FRS and PCE are modest with typical ranges of 0.66–0.76 and 0.68–0.76, respectively[4], leaving potential room for improvement. Thus, the discovery of additional biomarkers that improve or independently inform the predictive power of these existing models has been the objective of multiple recent research endeavors[5,6].

Imaging biomarkers derived from computed tomography (CT) have shown promise in the assessment of cardiovascular risk. For example, the coronary artery calcium (CAC) score measures the extent of plaque in the coronary arteries from coronary CTs, and is an important tool for IHD risk stratification[7,8]. Although CAC scoring is a strong independent predictor of cardiovascular events[9], the integration of both clinical factors (e.g., FRS) and imaging factors (e.g., CAC score) has been shown to significantly improve prediction of major cardiac events and all-cause mortality (compared with clinical or imaging metrics alone)[10,11]. Other studies have combined metrics from coronary CT angiography with blood biomarkers such as high-sensitivity cardiac troponin to successfully improve upon current measures of cardiovascular risk[12,13]. These specialized methods apply to a subset of patients already being assessed for cardiovascular risk.

Alternatively, abdominopelvic CTs contain body composition (BC) imaging biomarkers for atherosclerotic cardiovascular disease, such as hepatic steatosis[14], low muscle mass[15], an increased ratio of visceral to subcutaneous adipose tissue (VAT/SAT)[16], and abdominal aortic calcification[17]. Notably, 20 million abdominopelvic CTs are acquired annually almost twice as often as CT scans that image the heart or coronary vessels, such as non-contrasted chest CT and coronary CT[18,19]. According to the National Hospital Ambulatory Care Survey (https://bit.ly/2SL6957), in 2016 over 10 million abdominopelvic CTs were acquired in the US during emergency department visits alone, often in relation to abdominal pain—the most common principal reason for visiting an emergency department[20]. By comparison, roughly 3 million chest CTs were performed during emergency department visits in 2016. Within abdominopelvic CTs, these biomarkers could be measured during such routine imaging procedures without resulting in additional costs or radiation exposure, referred to as opportunistic imaging[21]. However, the current clinical workflow and volume of imaging is not well-suited to allow practical utilization of the additional resources required to manually extract measurements of imaging biomarkers[22]. Consequently, despite the potential value, cardiovascular risk is not routinely assessed upon abdominopelvic CT acquisition, thereby missing opportunities for early disease detection and prevention.

In this work, we developed IHD risk assessment models that use automatically measured imaging features from abdominopelvic CT examinations in combination with the patient's EMR. We evaluate the benefit of extracting BC imaging biomarkers from an axial slice at the level of the third lumbar vertebra (L3) in addition to traditional PCE metrics. We also develop an IHD risk assessment tool using the raw L3 slice image in an end-to-end manner using deep learning. We further develop a method to quantitatively assess the contribution of imaging features to the model prediction, aggregated at the tissue level. We introduce this method, Tissue Saliency, in this work. Finally, we combine features derived from the EMR in addition to the L3 slice, yielding the greatest risk prediction performance, and interpret the individual contribution of clinical features. To spur further research, we publicly release the Opportunistic L3 for IHD prediction (OL3I) dataset. Overall, we depict how opportunistic utilization of already-acquired CT imaging and EMR data can facilitate primary prevention of IHD without requiring additional testing, radiation, cost, or radiological assessment.

Methods

Study population

Following Stanford University Institutional Review Board approval and in accordance with relevant guidelines and regulations, we identified an initial cohort of 36,354 contrast-enhanced abdominopelvic CTs performed for abdominal pain between January 2013 and May 2018 on individuals who presented to our tertiary center emergency department. We included images with 1.0 or 1.25 mm axial spacing, from individuals 18 years of age or older with at least one documented clinical encounter in the year prior to and at least 1 year immediately following the acquisition of the image. For each patient, data from previous medical encounters were obtained. Informed consent was waived for this analysis by our Institutional Review Board. All demographic information (birthdate, sex, race/ethnicity), along with vital signs, body mass index (BMI), International Classification of Disease 10th edition (ICD10) codes, Current Procedural Terminology (CPT) codes, laboratory results, and prescriptions were extracted. We labeled individuals who had an ICD10 diagnosis code for Ischemic Heart Diseases (I20-I25) in the follow-up period after the image acquisition as IHD positive and those that did not have the code as negative. ICD codes have been found to have high sensitivity and specificity in identifying IHD in prior studies[23,24]. Since our goal was to identify new IHD patients that may not otherwise be detected, we excluded images from individuals with any diagnosis of IHD prior to and at the time of the image acquisition. We defined two cut-off periods for follow-up, 1 year and 5 years, establishing two cohorts representing individuals who either develop IHD or have follow-up within those time frames.

From each CT volume, we automatically identified the slice at L3 using a previously published convolutional neural network (CNN) algorithm[25], manually verifying correctness for each case. The L3 slice was chosen as it is the most common reference location for BC analysis[26–29]. The process to select the final cohorts of patients is shown in Supplementary Fig. 1. We excluded images with artifacts that obscured the L3 level (e.g., spinal

instrumentation), those that had anatomical variations in the image (e.g., scoliosis) that precluded the assignment of a single slice to the L3 level, those that did not contain the L3 slice in the field of view, and those obtained within the same 6-month window as an already included image. For each cohort, we used random sampling (stratifying on outcome labels) to divide patients in the dataset into training and test datasets representing 80% and 20% of the images, respectively, for IHD risk estimation and model creation.

Segmentation model

Given that BC metrics from manual segmentations have been correlated with cardiovascular risk, we trained a 2.5D U-Net CNN to perform BC analysis for segmenting regions of muscle along with VAT and SAT on an abdominopelvic CT slice[30]. A total of 400 axial L3 slices obtained exclusively from the training set and manually labeled were used during model tuning and evaluation. Manual segmentation of muscle, VAT, SAT, and bone were performed semi-automatically with CoreSlicer[31], a free online tool, using attenuation thresholds and manual adjustments as needed (AW, 5 years of experience). 320 images were randomly selected for segmentation training, 40 for validation, and 40 for testing. Segmentation performance on the 40 test images was determined using segmentation accuracy metrics, namely the Dice coefficient and root-mean-squared coefficient-of-variation. The Dice coefficient is calculated as two times the ratio of the intersection between the ground-truth and segmented image masks to the sum of the number of pixels in each mask. A Dice score of 1 indicates a perfect segmentation. The variations between the manual and automated segmentation approaches on the tissue-wise radiodensity (measured in Hounsfield units [HU]) and cross-sectional area (measured in cm^2) were also evaluated.

The inputs to the 2.5D network were individual 2D axial CT slices at the L3 level with three different window and level (W/L) settings for maximizing tissue contrast. The W/L settings that were used included a soft tissue window (W/L = 400/50 HU), a bone window (W/L = 1800/400 HU), and a custom window (W/L = 500/50 HU). After applying the appropriate windowing, each of the channels was normalized to values between 0 and 1.

The U-Net utilized 6 convolution levels (each with two convolutional operators, both followed by a rectified linear unit activation, followed by batch normalization) for the encoder and decoder[32]. The number of U-Net features per layer increased quadratically from 32 to 1024. The dimensions of the convolutional kernels were 3×3, while that of the maximum pooling operator was 2×2. A softmax activation was used as the final layer in the CNN along with a weighted soft Dice loss function to account for class imbalance amongst the segmented tissues. The U-Net hyperparameters had previously been optimized for medical imaging segmentation[32]. A weighting factor of 8 was used for muscle during loss computation. All network weights were randomly initialized using the He et al.[33] initialization scheme.

Training was performed with the Adam optimizer with default parameters β_1: 0.9, β_2: 0.999, with a learning rate schedule that included a base learning rate of 1e−3 and the learning rate being reduced by 0.8 for every epoch to a minimum value of 1e−8. The network was nominally trained for 130 epochs with an early stopping criterion of a minimum change in loss of 1e−5 and a patience of 8 epochs. The batch sizes for training, validation, and testing were chosen to be 10, 33, and 80, respectively, for maximizing GPU memory. Training was performed using a Tensorflow 1.14 on an NVIDIA Titan Xp GPU.

We used the segmentations generated by the U-Net model and determined average muscle radiodensity in HU and the VAT/SAT cross-sectional area ratio. We trained a model (L2 logistic regression) using tenfold cross validation to select the L2 penalty weight. The model was trained using the training sets to predict IHD outcome at 1 and 5 years using these two features. We refer to this as the *Segmentation Only* model.

Imaging only model

We trained a CNN as a feature extractor to predict the risk of IHD using a single axial slice at the L3 level, using an EfficientNet-B6 architecture[34]. EfficientNets were designed to balance the scaling of network width, depth, and image size, thus producing state-of-the-art results in conventional image classification tasks with smaller and faster models as compared to traditionally used feature extractors, such as ResNet50[35]. The 512×512 pixel grayscale L3 slice was clipped to contain values from − 1000 to 1000 HU, represented as an unsigned integer, replicated thrice to produce a $3 \times 512 \times 512$ image, and resized to $3 \times 528 \times 528$ to be input into the network. The initial EfficientNet-B6 model weights were obtained from a pre-trained model optimized for ImageNet classification performance (https://pypi.org/project/efficientnet-pytorch/)[36]. The final fully-connected layer was replaced with one corresponding to a binary outcome, and the model weights were fine tuned. The tuning of the weights was performed with a cross-entropy objective using a random selection of 80% of the training set, reserving 20% for validation. Training was performed for multiple epochs until no improvement in validation loss or Area Under the Receiver Operating Characteristic (AUROC) was observed. A batch size of 8 and an Adam optimizer[37] was used with default parameters β_1 0.9 β_2 of 0.999 and a constant learning rate of 7e-6 and 6e-6 for the 1 and 5-year cohorts, respectively. Model training was carried out using Pytorch 1.1 on an NVIDIA Titan Xp GPU.

We compared training only the final layer as opposed to training all of the model weights, using additional image augmentations such as rotations of up to 3° and pixel shifting of up to 5 pixels during training, and using a focal loss function assigning higher weights to IHD cases. We chose the final network architecture and training strategy described above as it achieved the highest AUROC in the validation stage (Supplementary Table 1).

Clinical only model

We used the data extracted from the EMR to produce features to develop a predictive model. Demographic data used were age at time of scan and sex. In this initial approach, we did not include race/ethnicity as a feature because of its limited accuracy in medical records[38], in addition to obtaining no benefit in discrimination performance in preliminary results when including it as a covariate. From the EMR within one year prior to image acquisition, we obtained blood pressure measurements, BMI, relevant laboratory results (total, low-density

lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, fasting glucose and hemoglobin A1c), and diagnosis (ICD10), procedure (CPT), and medication (ATC) codes. In multiple clinical encounters, vital signs and laboratory results were combined using an exponential weighting average, with each weight inversely proportional to the difference in time between the data point acquisition and the image acquisition. The number of times a vital sign or lab result was reported was also used as a feature. Apart from select PCE covariates (low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, imputed with median imputation), no imputation strategy was used for missing values.

To avoid sparsity, we grouped ICD10, CPT, and medications based on their underlying ontology. Namely, we grouped ICD10 codes (https://bioportal.bioontology.org/ontologies/ICD10) by blocks and CPT Category I codes (https://bioportal.bioontology.org/ontologies/CPT) at the H2 level. To further summarize overall burden of disease in a single feature, we also calculated the Charlson Comorbidity Index for each patient and included it as a feature[39]. Irrespective of dose and frequency, the active substance of prescribed medications was mapped to RxNorm and subsequently to the second level of the Anatomical Therapeutic Code (https://bioportal.bioontology.org/ontologies/ATC), corresponding to the therapeutic subgroup. Furthermore, prior to training, we iteratively removed highly correlated features (Pearson correlation coefficient > 0.5). In all, each patient EMR was represented using a 422-dimensional vector. The final clinical features used and their descriptions are listed in Supplementary Table 2.

The predictive model used for predicting IHD risk from EMR features was designed using XGBoost, an optimized gradient-boosting machine learning system[40]. In gradient boosting, an ensemble of weak learners is iteratively constructed by greedily adding estimators that fit the previous residual. In doing so, gradient boosting algorithms can perform successfully across a wide variety of predictive tasks, often outperforming traditional models such as logistic regression or support vector machines. We chose XGBoost for its robust performance in predictive modeling, and for its capacity to handle missing data, which other gradient boosted methods like AdaBoost lack. Optimal parameters for training the model were established using ten-fold cross-validation on the training set.

Fusion models

To further identify the potential benefits of using imaging BC features as predictors of IHD risk, we constructed three models to fuse imaging and clinical data. In the first fusion, we concatenated the features used by the PCE with the average muscle radiodensity and the VAT/SAT ratio (*PCE + Segmentation Model*), the latter two measurements obtained by using our automated segmentation model. We used an XGBoost classifier to predict IHD risk with the PCE + Segmentation features. In the second fusion, we combined the risk output from our EfficientNet-B6 model with the risk output from our medical record model using stacking with L2 logistic regression (*Imaging + Clinical Model*). In the third fusion, we combined the risk output from the *Imaging Only*, *Clinical Only,* and *Segmentation Only* models (*Imaging + Clinical + Segmentation Model*) in a stacking L2 logistic regression classifier. In all fusion cases, we performed a hyperparameter search in tenfold cross-validation in the training set. A summary of our prediction model approach can be seen in Fig. 1. The clinical model and fusion models were trained using scikit-learn 0.23 (https://scikit-learn.org/) in Python 3.6.

Interpretation of model performance

Two baseline models currently employed in clinical practice that estimate 10-year cardiovascular risk were used as a reference, namely the FRS[1] and the PCE[2]. We studied the performance of the FRS as it directly models the risk of hard IHD events. Despite the PCE including other atherosclerotic cardiovascular disease outcomes, we also included them as a baseline given their current use in clinical practice guidelines. Since several patients were missing covariates necessary for FRS and PCE calculation (Supplementary Table 3), these values were imputed using median imputation to allow for a baseline risk calculation for all individuals in the study. In addition, we examined the performance of all models in the subpopulations with available/missing PCE covariates (Supplementary Table 6).

Attribution analysis

With the aim of allowing for interpretability of the fusion model, we examined the contributions of individual features in both the imaging and clinical models.

For the imaging model, we developed a new tissue saliency interpretation tool as described in the introduction to evaluate the tissues that had a large contribution to the final prediction outcome. We first calculated the derivative, w, of the IHD class score at the final layer of our EfficientNet-B6 model, S_{IHD}, with respect to each input pixel, I_{ij}, in the image I[41]. That is,

$$w_{ij} = \left.\frac{\partial S_{\mathrm{IHD}}}{\partial I}\right|_{I_{ij}}.$$

Using our segmentation model, each pixel was assigned a specific tissue class, t. In our particular case, this corresponded to either muscle, VAT, SAT, other body tissues, or background. We obtain the observed normalized tissue saliency, S^O, for a particular tissue, as:

$$S_t^O = \frac{|w_{ij}(t)|}{|w_{ij}|}.$$

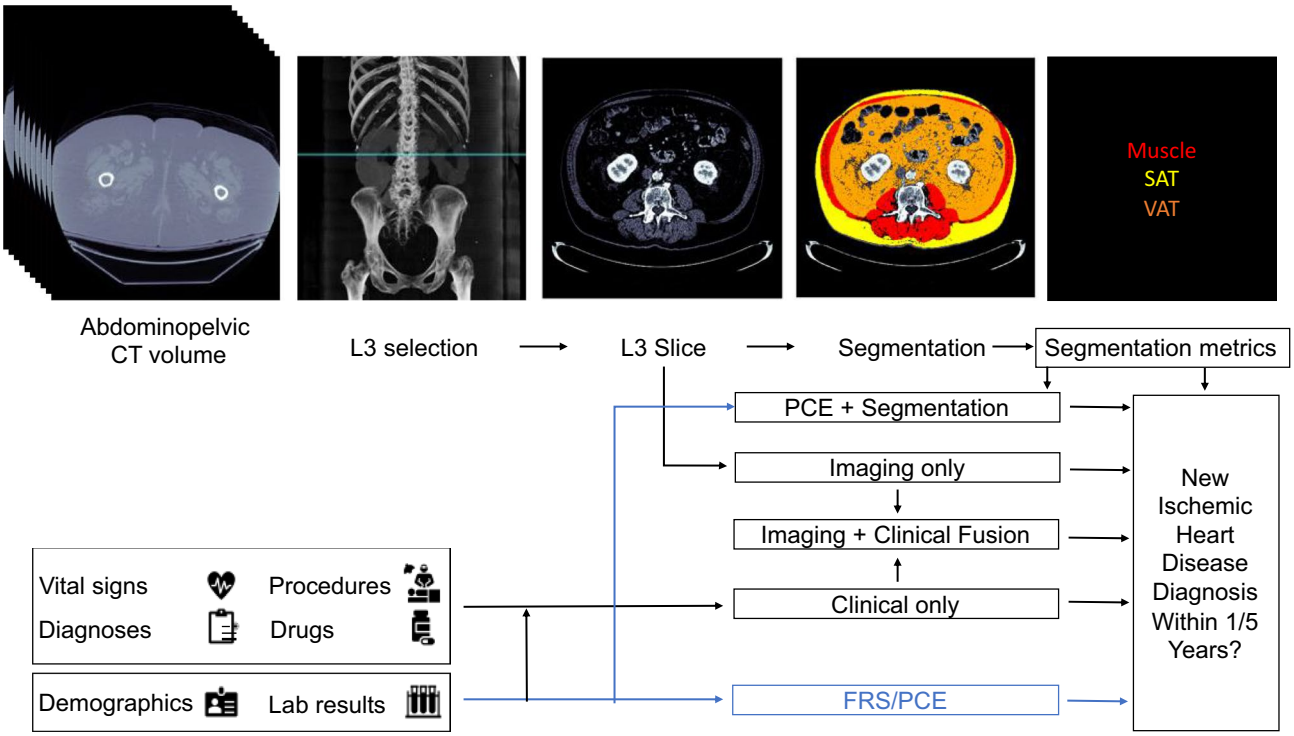

Figure 1. Proposed models for evaluating risk of a future ischemic heart disease diagnosis in one or five years following an abdominopelvic computed tomography (CT). The blue line shows which sources are used by existing risk assessment models, the Framingham Risk Score (FRS) and Pooled Cohort Equations (PCE). The PCE is the standard tool for atherosclerotic cardiovascular disease risk assessment in current clinical guidelines for 10-year risk estimation. In our proposed models, the axial slice corresponding to the third lumbar vertebra anatomical level (L3) is automatically selected from the CT volume. In one model, the L3 slice is automatically segmented to extract mean muscle radiodensity in Hounsfield units and the Visceral/Adipose cross-sectional area ratio; these features are used alone or in combination with covariates from the PCE to form a *Segmentation* or *PCE + Segmentation Model*. Alternatively, features are automatically extracted from the L3 slice using a convolutional neural network (*Imaging Only Model*). As an additional approach, predictions from a model trained with features constructed from the patient's electronic medical record within the year prior to CT acquisition (*Clinical Only Model*) are stacked with those of the imaging model (*Imaging + Clinical Fusion Model*) and with those of the *Segmentation* model (*Imaging + Clinical + Segmentation Model,* not depicted).

That is, the L1 norm of saliency of a particular tissue divided by the L1 norm of the total saliency. We contrast this observed saliency value with the expected tissue saliency, S^E, which we define using the cardinality of the observed segmentation:

$$S_t^E = \frac{|I_{ij}(t)|}{|I_{ij}|},$$

i.e., the proportion of pixels in the image belonging to the tissue t. We compared the proportion of observed vs. expected tissue saliency across the dataset by averaging the values for each image, $\overline{O}/\overline{E}$.

To assign specific tissue labels for each pixel, we first obtained a binary mask identifying the patient's body by removing the background and CT bed using traditional image processing methods[42]. Specific tissue labels for adipose tissue and muscle within the body were automatically assigned using our segmentation CNN. Tissues within the body mask not belonging to VAT, SAT, or muscle were assigned to the *Other Tissues* class. All other pixels were considered the background. Overall, our tissue saliency analysis depicts which tissue phenotype contributes most towards IHD risk.

To examine the relative contribution of each clinical feature to the prediction decision, we determined the Shapley Additive exPlanations (SHAP) values for each feature for each individual. SHAP values are an additive metric of feature importance that quantify the change in expected model prediction conditioned on a feature value[43]. In other words, they are a measure of how much the model prediction changes given a value for a particular feature. We used SHAP values to interpret the relative contribution of our clinical features to the final model classification output.

Statistical analysis

The AUROC was used as the primary metric to compare and select models during training. The AUROC is equivalent to the c-statistic in the case of binary classification, as is traditionally reported in the cardiovascular

risk assessment literature. In addition, we measured the Area Under the Precision Recall Curve (AUCPR), which is more informative in the case of imbalanced datasets, such as this one[44]. For both metrics, 95% confidence intervals were obtained using the stratified bootstrap method. We also report the sensitivity, specificity, and positive and negative predictive values of our models at a sample threshold defined using Youden's index.

Statistical analyses were carried out using SciPy 1.3[45]. Comparisons between AUROC values were carried out using the DeLong method, as has been previously validated for this purpose[46]. Comparisons between AUCPR values were established using the stratified bootstrap method. Observed and expected tissue saliency values were compared using paired t-tests. All tests performed were two-tailed. An α value of 0.05 was used to determine statistical significance.

Results

Final patient cohort

We collected a dataset of 8139 CT images of individuals with at least 1-year of follow-up, with a subset of 1747 images of 1671 individuals with at least 5-years of follow-up. The average (interquartile range) length of follow-up was 3.6 (2.2) years. For each individual, data available in the EMR in the year before the scan acquisition was obtained. With 1- and 5-year follow-up after CT scan acquisition, a new IHD diagnosis was identified in 355 (4.4%) and 440 (25.2%) of individuals. Because sampling was performed in a stratified fashion, the prevalence in the training and test sets was equal for both cohorts. The average (standard deviation) age at time of scan in the dataset was 51.7 (17.1) years, with 40.5% of CT exams in men. Additional demographic characteristics of both cohorts, along with PCE covariates and BC metrics, are in Supplementary Table 3.

CT scans were performed on 14 multi-slice CT scanners from GE and Toshiba (n = 4643 and 3496), respectively. Parameters for CT protocols included a tube voltage mode of 120 kV (range 70–140 kV), slice thickness of 1.00 or 1·25 mm (n = 3496 and 4643 respectively), and tube current setting based on BMI (mean 424.5 mA, standard deviation 179 mA). Soft tissue reconstruction kernels were predominantly used. Additional details on CT scanners and protocols used are listed in Supplementary Table 4.

The L3 slice was correctly labeled in 8098 (99.5%) cases. In the 41 incorrect cases, the predicted L3 slice was 137 ± 126 mm away from the correct location (mean ± standard deviation). Incorrect localization typically occurred on CT exams with additional anatomical coverage, such as scans also including the chest or lower extremities. Otherwise, the automatically selected slice was at the L2 or L4 level, immediately adjacent to the L3 level.

The performance of the models on the held-out test set is reported as follows.

Traditional IHD risk assessment model performance

The PCE performed comparably to the FRS in 1-year IHD prediction (AUROC 0.77 vs 0.74; $P = 0.07$; AUCPR 0.13 vs 0.10; $P = 0.05$), and 5-year IHD prediction, with AUROC 0.73 vs 0.71 ($P = 0.14$) and AUCPR 0.44 vs 0.43 ($P = 0.79$) respectively (Table 1). The ROC and precision-recall curves for the PCE are shown in Fig. 2 and compared with the FRS and other proposed models in Supplementary Fig. 2. Sensitivity, specificity, and positive and negative predictive values for PCE at two clinically relevant cut-offs are shown in Supplementary Table 5.

Segmentation model performance

Example segmentations produced by the model are depicted in Fig. 3a. These examples, along with a quantitative assessment of the model as measured by the Dice scores (Supplementary Fig. 3), show that the model can reliably label the muscle and adipose tissue, with a mean (standard deviation) Dice score of 0.97 (0.03), 0.97 (0.02) and 0.96 (0.05) for muscle, SAT and VAT, respectively. Furthermore, the error in computing tissue radiodensity and cross-sectional area was below 1% and 2%, respectively, for the three segmented tissues.

The bi-variate predictive model using VAT/SAT ratio and L3 muscle radiodensity as features (*Segmentation Only Model*) performed inferior to the PCE for 1-year IHD risk estimation, with AUROC of 0.70 ($P = 0.02$) and

Model	1y AUROC (95% CI)	*P*	1y AUCPR (95% CI)	*P*	5y AUROC (95% CI)	*P*	1y AUCPR (95% CI)	*P*
FRS	0.74 (0.68–0.79)	0.07	0.10 (0.08–0.13)	0.05	0.71 (0.65–0.77)	0.14	0.43 (0.36–0.52)	0.79
PCE	0.77 (0.71–0.81)	–	0.13 (0.10–0.19)	–	0.73 (0.68–0.79)	–	0.44 (0.38–0.54)	–
Segmentation	0.70 (0.65–0.76)	0.02	0.08 (0.07–0.12)	0.04	0.68 (0.62–0.75)	0.10	0.36 (0.31–0.44)	0.06
PCE + segmentation	0.76 (0.71–0.81)	0.83	0.11 (0.09–0.14)	0.38	0.72 (0.66–0.78)	0.40	0.41 (0.35–0.51)	0.39
Clinical only	0.80 (0.75–0.84)	0.08	0.14 (0.11–0.21)	0.70	0.76 (0.71–0.81)	0.27	0.51 (0.43–0.61)	0.14
Imaging only	0.76 (0.70–0.81)	0.89	0.12 (0.09–0.19)	0.76	0.78 (0.71–0.83)	0.25	0.56 (0.47–0.66)	0.048
Imaging + clinical fusion	0.81 (0.76–0.85)	0.02	0.15 (0.12–0.22)	0.49	0.80 (0.74–0.85)	0.03	0.60 (0.51–0.70)	0.003
Imaging + clinical + segmentation fusion	0.80 (0.75–0.84)	0.14	0.15 (0.11–0.22)	0.53	0.81 (0.75–0.86)	0.02	0.63 (0.54–0.72)	< 0.001

Table 1. Proposed model performance in comparison to pooled cohort equations (PCE) and Framingham risk score (FRS) as measured by area under receiver operating characteristic (AUROC) and precision-recall (AUCPR) curves. 95% confidence intervals (CI) and *P* values were obtained using the DeLong method for AUROC and the bootstrap method for AUCPR. Reported *P* values correspond to comparisons with the PCE.

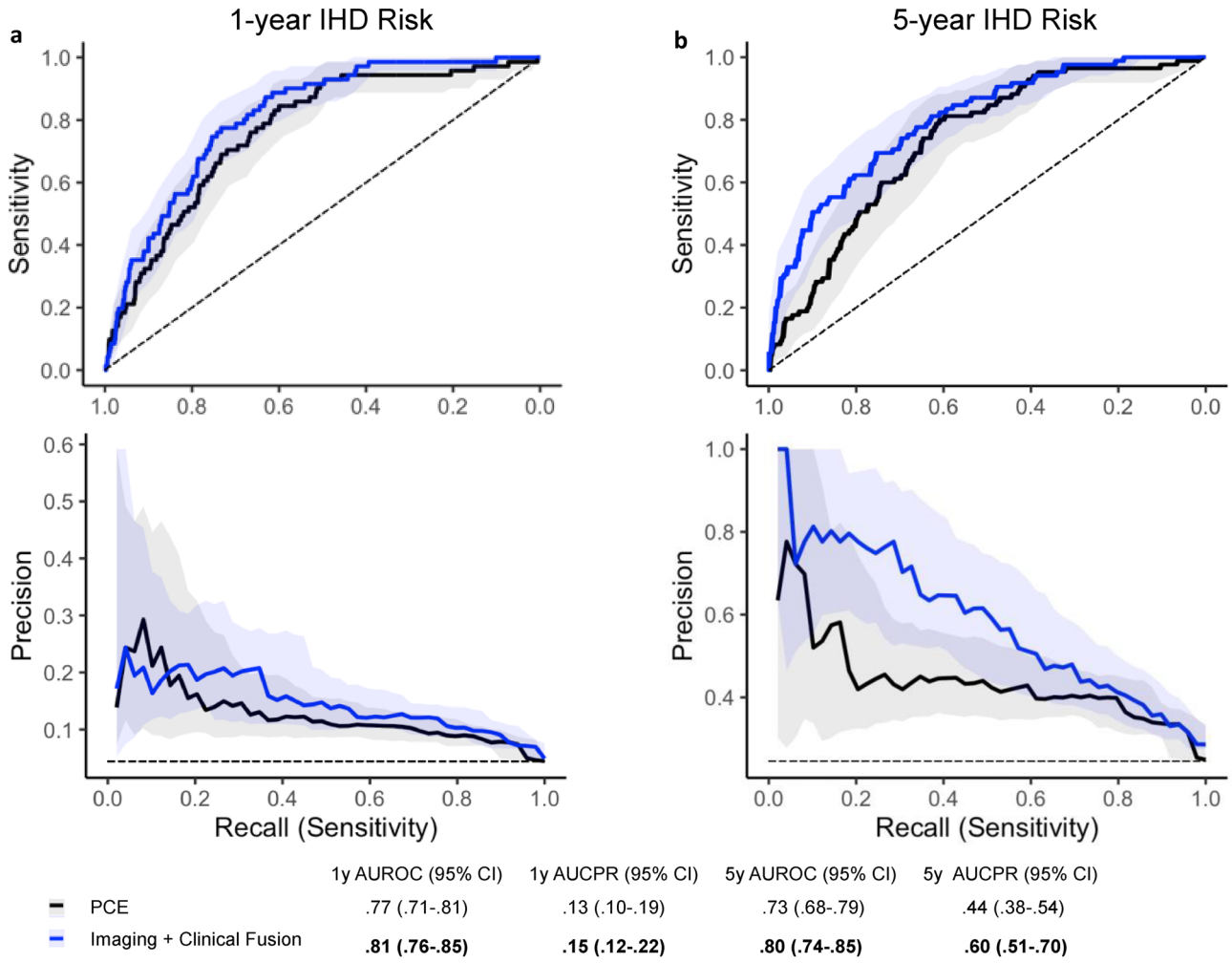

Figure 2. Performance of proposed Imaging + Clinical Fusion model compared to Pooled Cohort Equations (PCE), visualized through the receiver operating characteristic (ROC) curves (top row) and precision recall curves (bottom row) for (**a**) 1-year and (**b**) 5-year IHD risk modeling. Shaded lines indicate 95% confidence intervals (CI). Dashed lines show performance of a random (ROC curve) or a prevalence-based classifier (precision recall curve) as the simplest baselines. Area under the curve (AUC), 95% CI were determined using DeLong's method for the ROC curve and using the bootstrap method for the precision-recall curve.

AUCPR of 0.08 ($P=0.04$). In the 5-year cohort, the model performed comparably, with AUROC of 0.68 ($P=0.10$) and AUCPR of 0.36 ($P=0.06$).

Imaging only model performance

The *Imaging Only Model* also achieved comparable performance to the PCE for 1-year IHD risk prediction. It achieved a 1-year AUROC of 0.76 ($P=0.89$) and AUCPR of 0.11 ($P=0.76$). It showed improved performance in the 5-year risk prediction, with comparable AUROC (0.78; $P=0.25$), yet higher AUCPR (0.56; $P=0.048$) (Table 1). This model also outperformed the *Segmentation only* model, with statistically significant increases of 0.06 ($P=0.04$) and 0.10 ($P=0.005$) in 1 and 5-year AUROC, respectively.

The contribution of individual tissues to the final prediction was assessed through tissue saliency. Figure 3a shows sample L3 slice segmentations, as well as tissue saliency values superimposed on the original image. Figure 3b shows the distribution of observed and expected tissue saliency values. For the 1-year follow-up cohort, the observed/expected tissue saliency ratios were 1.71, 1.67, 1.50, 1.15, and 0.54 for Other Tissues, muscle, VAT, SAT, and background, respectively. For the 5-year follow-up cohort, the ratios were 2.00, 1.76, 1.75, 1.50, and 0.42 for VAT, Other Tissues, muscle, SAT, and background, respectively. That is, these ratios were higher than expected for muscle, VAT, SAT, and other body tissues, and lower than expected for the background. All differences in pairs of observed vs. expected values were statistically significant ($P<0.001$).

Clinical only model performance

The *Clinical Only* model achieved AUROC/AUCPR of 0.80/0.14 and 0.76/0.51 for 1 and 5-year IHD risk prediction, achieving comparable performance to the PCE in 1-year prediction (AUROC $P=0.08$, AUCPR $P=0.70$) and in 5-year prediction ($P=0.27$ for AUROC and 0.14 for AUCPR).

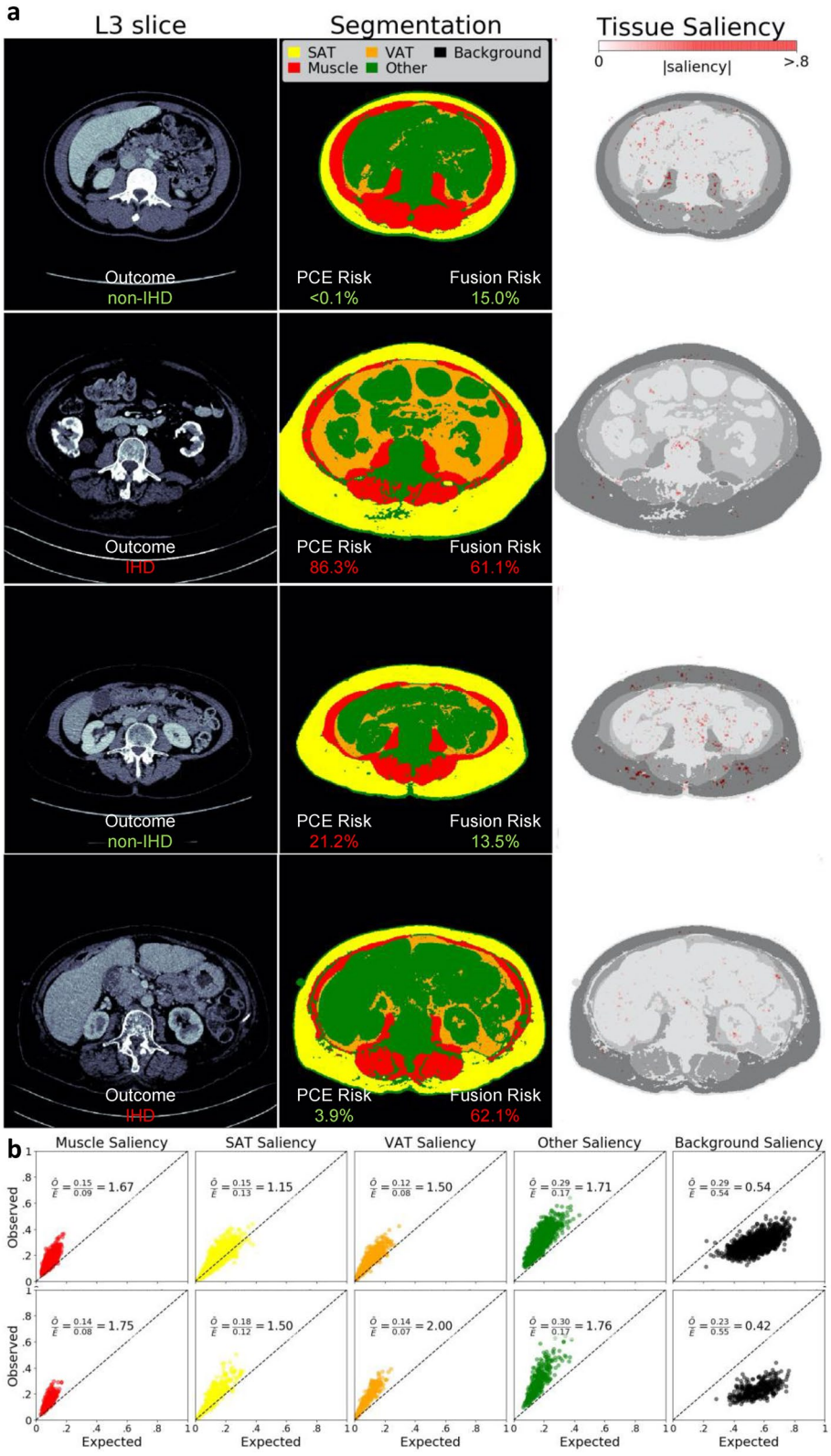

Figure 3. Segmentations and tissue saliency. In (**a**), sample L3 slices (first column) are shown for four individuals with at least 5-years follow-up after their image was acquired, with their corresponding segmentations generated from the segmentation model (second column). Their calculated risk from the traditional PCE is contrasted with the more accurate Imaging + Clinical fusion risk. The saliency from the imaging model is shown overlayed on the segmentation (third column). In (**b**) the distributions of observed (aggregated saliency values for each tissue type relative to the saliency for the image) versus expected saliency are shown for each tissue, for 1-year (top) and 5-year (bottom) risk prediction, where expected saliency is calculated as the proportion of pixels corresponding to a class in the image.

Figure 4a shows the ten features with the highest average SHAP value in the 5-year IHD risk prediction model. The majority of these top 10 features were also identified in the 1-year follow-up cohort, albeit not in the same order. In both cases, traditional cardiovascular risk factors, such as age, male sex, and hypertension-related variables, are present among the top features. In addition, serum glucose measurements and diagnosis of renal failure were present among the top 10 for both models. The top features for the 1-year model included the number of times serum glucose was measured, the use of diuretics, symptoms, and signs involving the circulatory and respiratory system and BMI, which were not present among the top ten for the 5-year model. The most impactful feature on prediction as determined by SHAP value was age, with higher risk for older individuals. SHAP values for 4 individuals from the 5-year risk cohort are shown in Fig. 4b.

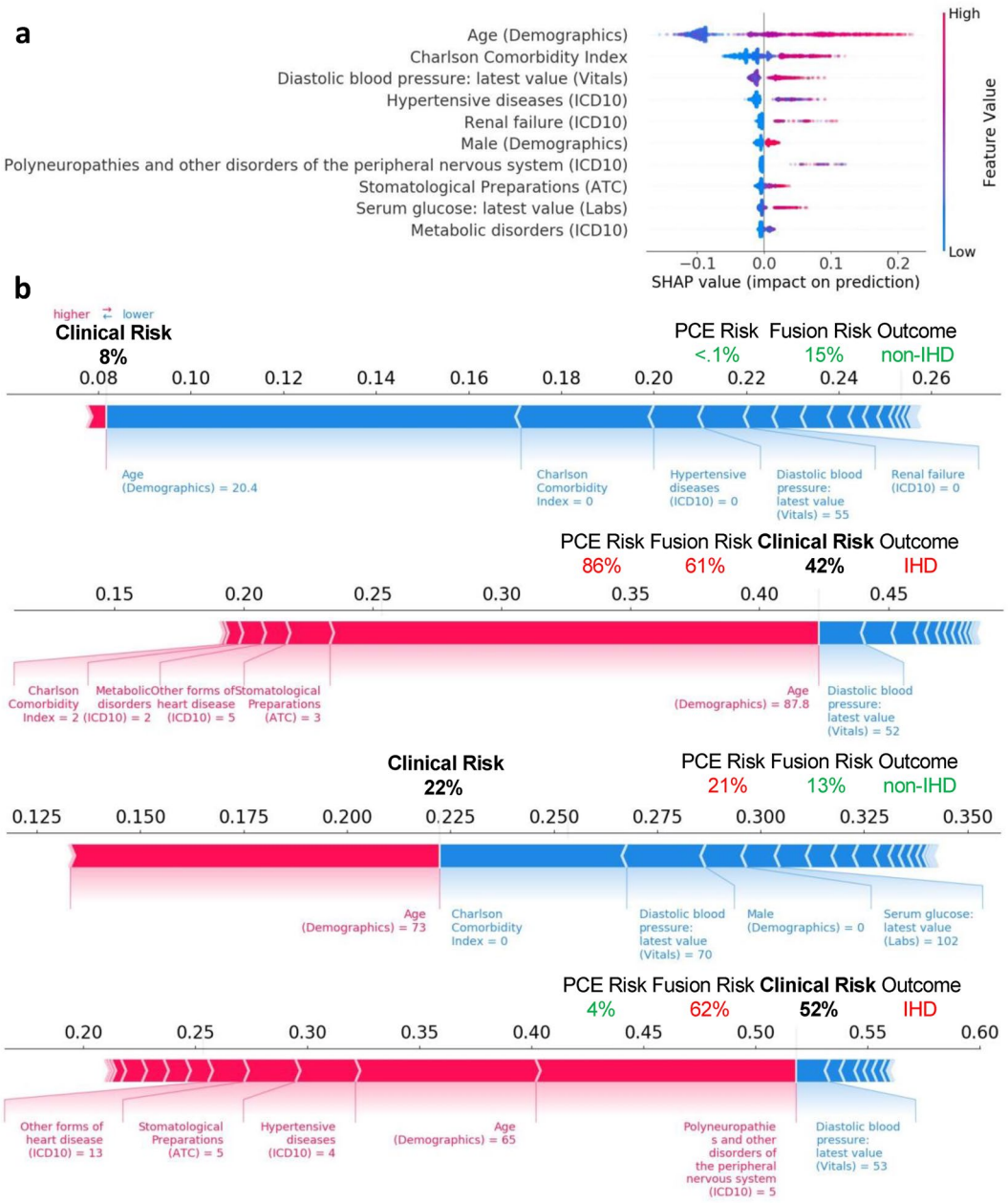

Figure 4. *Clinical Only* model feature importance as quantified by SHAP (SHapley Additive exPlanations) values for the top 10 features in the training set of the 5-year risk prediction cohort (**a**). Higher SHAP values indicate higher than expected probability of IHD as assigned by the model. Individual SHAP values for features with highest values for 4 individuals from the 5-year risk prediction cohort (**b**), along with the risk assigned by the *Clinical Only* model. Their PCE and Imaging + Clinical Fusion model risk are also shown, along with the outcome.

Fusion model performance

The performance of the predictive algorithms combining the BC segmentation metrics with the original PCE covariates (*PCE + Segmentation Model*) was comparable to that of the PCE alone, with 1 and 5-year AUROC/AUCPR values of 0.76 ($P = 0.83$)/0.11 ($P = 0.38$) and 0.72 ($P = 0.40$)/0.41 ($P = 0.39$), respectively. Thus, inclusion of average muscle radiodensity and VAT/SAT ratio did not improve prediction of IHD risk using PCE covariates.

The performance of the *Imaging + Clinical Model* is depicted in Fig. 2. For 5-year IHD risk modeling, this model showed marked improvement in prediction capabilities compared to the PCE, both in terms of AUROC (0.80) and AUCPR (0.60), with a statistically significant increase ($P = 0.03$, $P = 0.003$, respectively) of 0.07 and 0.16, respectively. For 1-year IHD risk modeling, the model had higher AUROC compared to the PCE (0.81, $P = 0.02$) with similar AUCPR.

The *Imaging + Clinical + Segmentation Model* performed similarly to the *Imaging + Clinical Model* for 1-year IHD prediction. No statistically significant differences were found in AUROC or AUCPR for 1 IHD prediction compared to the *Imaging + Clinical Model*. There was a statistically significant improvement of 0.03 in AUCPR ($P = 0.02$) compared to the *Imaging + Clinical Model* for 5-year IHD prediction, and a non-statistically significant improvement of 0.01 AUROC ($P = 0.25$). Furthermore, a statistically significant increase in AUROC and AUCPR for 5-year IHD prediction compared to the PCE, as is the case of the *Imaging + Clinical Model,* was found.

The performance of baseline models and all proposed models across subpopulations with complete/missing PCE covariates, and across different age, sex, race/ethnicity, among patients taking lipid-modifying agents, as well as those with acute vs. nonacute IHD outcomes is reported in Supplementary Table 6.

Discussion

In this study we propose automated, explainable, and opportunistic risk assessment methods for 1-year and 5-year IHD risk following a contrast-enhanced abdominopelvic CT. Using a single slice from the CT, features quantifying the muscle radiodensity and body fat in conjunction with traditional cardiovascular risk factors and additional clinical data derived from the EMR, our models perform comparably or better than currently used tools to assess cardiovascular risk. Additionally, we publicly release all L3 images, corresponding EHR-derived features, and 1 and 5-year IHD outcomes (the OL3I dataset), which is the first large-scale public dataset for opportunistic CT evaluation with patient outcomes. Our code and trained models are also made publicly available.

The use of BC biomarkers derived from abdominal CT imaging for cardiovascular risk assessment has been explored in the past. Pickhardt et al. extracted univariate metrics from CT colonography such as liver and muscle radiodensity, abdominal aortic calcification, and VAT/SAT ratio combined with FRS in asymptomatic individuals and found an improvement in 2-year cardiovascular event prediction AUROC of 0.77 compared to 0.71 using FRS alone[47]. While highlighting the value of imaging biomarkers in cardiovascular risk assessment, their methods were developed using CT colonography, an imaging modality that remains underutilized[48]. In contrast, our models perform opportunistic risk assessment in individuals that undergo contrast-enhanced abdominopelvic CTs, a more commonly used diagnostic scan in a wide variety of clinical settings. Thus, our model may potentially allow more opportunities for incidental risk assessment for IHD. Moreover, current diagnostic scans, such as abdominopelvic CTs, are generally geared towards addressing a primary clinical concern (e.g., cause for acute abdominal pain). Models such as the ones proposed in this study could increase the diagnostic and prognostic value of medical images by providing risk assessment in addition to answering the primary clinical question such as the etiology of a patient's acute symptoms. Magudia et al. predicted myocardial infarction using population-normalized BC metrics (muscle, VAT and SAT area z-scores) in outpatient adults without a major cardiovascular or oncologic diagnosis undergoing routine abdominal CT[49]. They found after controlling for BMI and other cardiovascular risk factors, only VAT area was associated with subsequent infarction risk. Though their approach is limited by their demographics (only White and Black individuals were included, with 89% of patients being White), these findings are consistent with our *PCE + Segmentation* models, which suggest that aggregate BC metrics alone may not be sufficient to improve predictions of clinically relevant baselines. Alternatively, our models include individuals of Asian and Hispanic race and ethnicity, as well as improved feature extraction of both imaging biomarkers and clinical features that result in improved predictive performance over existing clinical models.

Both the FRS and PCE have been shown to overestimate the risk of developing cardiovascular disease in contemporary, real-world populations[4]. As models used predominantly for time-to-event risk assessment, they have been developed to maximize the c-statistic (or AUROC in binary classification settings) and are typically used with cut-offs defined to have high sensitivity, at the expense of specificity. In our test cohorts, the AUROC of these baseline models was comparable to prior validation studies[4,50]. We believe that the use and reporting of AUCPR should be considered in the development of IHD risk assessment models, as it has been shown that (1) a curve will dominate in ROC if and only if it dominates in precision-recall space, and (2) PRC are more informative in an imbalanced classification setting, as is typical for IHD risk assessment[51]. By considering the trade-off between precision and recall, models can be designed to have a high sensitivity but also have a high precision (i.e., fraction of true positives among those identified as positive). Although our intent is not to replace their role in current risk assessment, we note that our 5-year IHD risk models outperform the PCE in both the ROC and precision-recall space, which indicates that they can successfully identify individuals at risk of developing IHD, with a higher proportion of true positives among those identified with high risk.

Our methods seek to address model interpretability, an important barrier for potential implementation of artificial intelligence in medicine[52]. Though our segmentation model had high Dice scores, similar to other published studies[53,54], BC metrics alone or in combination with PCE covariates did not outperform the PCE. To our knowledge, our approach is the first to model IHD risk prediction in an end-to-end manner using images

directly, as opposed to BC metrics, outperforming the radiomics/PCE approach in 5-year IHD prediction. In addition to treating IHD risk prediction as an end-to-end imaging problem, we introduced the concept of tissue saliency to study the contribution of pixels to the predictions made by the *Imaging Only Model*. Unlike existing widely used pixel-wise attribution methods that only enable qualitative assessments of individual images, tissue saliency enabled us to assess the contribution of groups of pixels belonging to the same tissue class in a qualitative and quantitative manner. As expected, tissues within the body had a higher amount of saliency than expected, most notably the VAT and muscle tissues. This is consistent with the observation that biomarkers quantifying the radiodensity or area of these tissues are informative of IHD risk[15]. The tissue saliency ratio for *Other Tissues* was also higher than expected, which could be due to the liver, abdominal aortic calcifications, or trabecular bone radiodensity being present in slices at this level, but not explicitly segmented in this study (as can be seen in examples of Fig. 3a). Furthermore, the background pixels provided a lower-than-expected but not negligible proportion of the saliency. Upon visualization of examples, background pixels with higher saliency are typically neighboring the body, which indicates that the model may be using patient habitus as a feature. In aggregate, tissue saliency provides insights into the contribution of tissues in the prediction, increasing our understanding of the underlying drivers of prediction in the better-performing *Imaging Only Model*.

We found evidence that including additional clinical features could improve the performance of the *Imaging Only Model*. Such EHR derived features have been leveraged to develop improved IHD risk assessment models compared to PCE[55]. Our results show the added value in combining them with imaging-derived features. We examined the importance of individual clinical features through their SHAP values. Among the top predictors were features that represent traditional cardiovascular risk factors, such as age, male sex, and hypertension. The Charlson Comorbidity Index had high importance in both models; previously, it has been correlated with recurrence and mortality following acute coronary syndromes[56,57], as well as anatomic severity of coronary artery disease[58]. Features such as increased serum glucose, which is associated with a higher risk in our models, may better assess IHD risk than the diagnosis of diabetes alone, a feature that was not salient among the SHAP analysis. Similarly, the use of diuretics, a common treatment for hypertension, was positively associated with IHD. This is explainable in that hypertension is a well-studied modifiable IHD risk factor[59]. Finally, renal failure (also among the top ten predictors) has been identified as an independent risk factor for IHD[60]. The use of SHAP values aids in understanding how an individual feature may affect the prediction of IHD risk. The presence of well-known risk factors among the top predictors raises trust in an individual attempting to scrutinize the prediction model. SHAP values and tissue saliency could provide clinicians with a mechanism to interpret and intervene based on specific aspects of the patient history. As new artificial intelligence applications in medical imaging continue to emerge and gain popularity, explainability may be an important factor for clinical adoption[61,62].

There are several potential ways our models could be utilized in clinical practice. After undergoing a commonly performed contrast-enhanced abdominopelvic CT for non-IHD indications, individuals could be opportunistically assessed for high IHD risk and undergo further cardiovascular follow-up or be referred to primary care or preventive cardiology for potential initiation of preventive interventions. Furthermore, the specific BC metrics automatically calculated from the CT could be used to identify tangible areas of improvement, as well as be used to track progress following an intervention. Ultimately, these models could identify an individual at high IHD risk that may have gone otherwise unnoticed, which is the goal of opportunistic risk assessment. Finally, the models could analyze CT scans that have already been performed and are housed within a picture archiving and communication system for retrospective identification of high IHD risk patients.

We publicly release the OL3I dataset, comprised of 8139 L3 images, accompanying clinical features, and IHD outcome labels. This is the first publicly available dataset for opportunistic imaging, and to the best of our knowledge the first multimodal CT dataset with prognostic outcomes. We expect this dataset will further promote research for imaging and multimodal approaches to opportunistic risk assessment of IHD.

Our study has limitations. Our data were sourced retrospectively from a single center. IHD diagnoses made outside the center are missed. Though our reported results correspond to a test set of patients that was not used during model development, confirming the potential clinical use of the models requires a prospective evaluation, ideally in multiple centers. This work lays the groundwork for such external evaluations, in which the applicability of our approach in underrepresented populations with differing race or socioeconomic backgrounds could be further analyzed. Though our cohorts were comprised of diverse individuals both in terms of sex and ethnicity (Supplementary Table 3), these have been identified as variables across which model performance may vary, typically to the disadvantage of underrepresented minorities[63]. We found small variations to be present within patient subpopulations (Supplementary Table 6), motivating further studies in other validation cohorts to identify demographic-specific thresholds for intervention[64]. Furthermore, we did not include socioeconomic indicators in our subanalyses, a factor of variation that has been previously detected in cardiovascular risk assessment[65]. Such evaluations could also address a potential selection bias in our study, where we selected patients that present to the emergency department with abdominal pain and have a contrast-enhanced abdominopelvic CT scan performed. We did not include additional body composition metrics such as waist circumference, waist:hip ratio or weight:height ratio, which may add predictive power to traditional risk scores, as they were not routinely measured in our patient population. Another limitation is that biomarkers such as aortic calcifications or liver radiodensity may not necessarily be visualized in the single L3 slice approach that we analyzed. This may explain the lack of improvement when combining segmentation metrics with PCE features that has been identified in other studies[17]. Additionally, we utilized a 1- and 5-year prediction time point which might penalize FRS and PCE, which were validated for 10-year risk assessment. However, 1-year prediction can enable the identification of very high-risk individuals[66], and other studies[67–70] have similarly used 5-year windows for model comparison. Additionally, we did not study ASCVD risk, which the PCE were developed for. The results presented herein are not meant to replace current risk assessment tools but rather serve as additional tools for assessment to aid clinical decision making. Finally, our 5-year cohort prediction models showed an improvement

over established baselines, but our 1-year IHD risk prediction models were not able to outperform them, despite exploring a variety of data sampling and alternative loss functions to improve the performance of these models. This improved performance in 5-year prediction, however, could provide a window of opportunity to initiate preventive interventions. Alternatively, a considerable proportion of individuals in our cohort were missing laboratory measurements that would be necessary to evaluate the PCE or FRS at the time of scan acquisition, indicating their cardiovascular risk was not recently assessed. Our opportunistic approach could be used to alert a referring provider of particularly high IHD risk individuals and prompt further evaluation.

In conclusion, we developed and open-sourced a framework to use artificial intelligence models that enable opportunistic risk assessment for IHD following an abdominopelvic CT scan. By drawing from multiple data sources, we were able to produce models that can perform comparably or better than currently used clinical risk models, which currently guide treatment decisions for individuals undergoing cardiovascular risk assessment. Models automatically integrating existing EMR and CT data to identify patients at increased risk for IHD could provide opportunities for more effective preventive interventions at a population health level.

Data availability

The OL3I dataset is publicly available at https://stanfordaimi.azurewebsites.net/datasets/3263e34a-252e-460f-8f63-d585a9bfecfc.

Code availability

To facilitate replication of these results and further research, our code and trained models are made publicly available at https://github.com/stanfordmimi/abct_ihd. Our trained segmentation models are available at https://github.com/StanfordMIMI/Comp2Comp.

Received: 26 September 2023; Accepted: 20 November 2023
Published online: 29 November 2023

References

1. Wilson, P. W. F. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
2. Goff, D. C. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *Circulation* **129**, 98. https://doi.org/10.1161/01.cir.0000437741.48606.98 (2014).
3. Arnett, D. K. *et al.* 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **140**, e596–e646. https://doi.org/10.1161/CIR.0000000000000678 (2019).
4. Damen, J. A. *et al.* Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: A systematic review and meta-analysis. *BMC Med.* **17**, 109. https://doi.org/10.1186/s12916-019-1340-7 (2019).
5. Serra, R., Ielapi, N., Barbetta, A., Andreucci, M. & De Franciscis, S. Novel biomarkers for cardiovascular risk. *Biomark. Med.* **12**, 1015–1024. https://doi.org/10.2217/bmm-2018-0056 (2018).
6. Choi, S. The potential role of biomarkers associated with ASCVD risk: Risk-enhancing biomarkers. *J. Lipid Atheroscler.* **8**, 173 (2019).
7. Kapoor, K., Cainzos-Achirica, M. & Nasir, K. The evolving role of coronary artery calcium in preventive cardiology 30 years after the Agatston score. *Curr. Opin. Cardiol.* **35**, 500–507 (2020).
8. Lee, H. *et al.* Machine learning and coronary artery calcium scoring. *Curr. Cardiol. Rep.* **22**, 1–6. https://doi.org/10.1007/s11886-020-01337-7 (2020).
9. Ayoub, C. *et al.* Prognostic value of segment involvement score compared to other measures of coronary atherosclerosis by computed tomography: A systematic review and meta-analysis. *J. Cardiovasc. Comput. Tomogr.* **11**, 258–267 (2017).
10. Commandeur, F. *et al.* Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: A prospective study. *Cardiovasc. Res.* **116**, 2216–2225 (2020).
11. Motwani, M. *et al.* Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis. *Eur. Heart J.* **38**, 500–507 (2017).
12. Gore, M. O. *et al.* Combining biomarkers and imaging for short-term assessment of cardiovascular disease risk in apparently healthy adults. *J. Am. Heart Assoc.* **9**, e015410 (2020).
13. Sandoval, Y. *et al.* Atherosclerotic cardiovascular disease risk stratification based on measurements of troponin and coronary artery calcium. *J. Am. Coll. Cardiol.* **76**, 357–370 (2020).
14. Mellinger, J. L. *et al.* Hepatic steatosis and cardiovascular disease outcomes: An analysis of the Framingham Heart Study. *J. Hepatol.* **63**, 470–476 (2015).
15. Kim, J. H., Cho, J. J. & Park, Y. S. Relationship between sarcopenic obesity and cardiovascular disease risk as estimated by the Framingham Risk Score. *J. Korean Med. Sci.* **30**, 264–271 (2015).
16. Kaess, B. M. *et al.* The ratio of visceral to subcutaneous fat, a metric of body fat distribution, is a unique correlate of cardiometabolic risk. *Diabetologia* **55**, 2622–2630 (2012).
17. O'Connor, S. D., Graffy, P. M., Zea, R. & Pickhardt, P. J. Does nonenhanced CT-based quantification of abdominal aortic calcification outperform the framingham risk score in predicting cardiovascular events in asymptomatic adults? *Radiology* **290**, 108–115 (2019).
18. 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).
19. Stopsack, K. H. & Cerhan, J. R. Cumulative doses of ionizing radiation from computed tomography: A population-based study. *Mayo Clin. Proc.* **94**, 2011–2021 (2019).
20. Larson, D. B., Johnson, L. W., Schnell, B. M., Salisbury, S. R. & Forman, H. P. National trends in CT use in the emergency department: 1995–2007. *Radiology* **258**, 164–173 (2011).
21. Boutin, R. D. & Lenchik, L. Value-added opportunistic CT: Insights into osteoporosis and sarcopenia. *Am. J. Roentgenol.* **215**, 582–594 (2020).
22. Parikh, J. R., Wolfman, D., Bender, C. E. & Arleo, E. Radiologist burnout according to surveyed radiology practice leaders. *J. Am. Coll. Radiol.* **17**, 78–81 (2020).
23. Kokotailo, R. A. & Hill, M. D. Coding of stroke and stroke risk factors using International Classification of Diseases, revisions 9 and 10. *Stroke* **36**, 1776–1781. https://doi.org/10.1161/01.STR.0000174293.17959.a1 (2005).

24. Cozzolino, F. *et al.* A diagnostic accuracy study validating cardiovascular ICD-9-CM codes in healthcare administrative databases. The Umbria data-value project. *PLoS ONE* **14**, e0218919 (2019).
25. Kanavati, F., Islam, S., Aboagye, E. O. & Rockall, A. Automatic L3 slice detection in 3D CT images using fully-convolutional networks. http://arxiv.org/abs/1811.09244 (2018).
26. Amini, B., Boyle, S. P., Boutin, R. D. & Lenchik, L. Approaches to assessment of muscle mass and myosteatosis on computed tomography: A systematic review. *J. Gerontol. Ser. A Biol. Sci. Med. Sci.* **74**, 1671–1678. https://doi.org/10.1093/gerona/glz034 (2019).
27. Derstine, B. A. *et al.* Quantifying sarcopenia reference values using lumbar and thoracic muscle areas in a healthy population. *J. Nutr. Health Aging* **22**, 180–185 (2018).
28. Wang, S. *et al.* The value of L3 skeletal muscle index in evaluating preoperative nutritional risk and long-term prognosis in colorectal cancer patients. *Sci. Rep.* **10**, 1–11 (2020).
29. van der Werf, A. *et al.* Percentiles for skeletal muscle index, area and radiation attenuation based on computed tomography imaging in a healthy Caucasian population. *Eur. J. Clin. Nutr.* **72**, 288–296 (2018).
30. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* 234–241 (2015).
31. Mullie, L. & Afilalo, J. CoreSlicer: A web toolkit for analytic morphomics. *BMC Med. Imaging* **19**, 15 (2019).
32. Desai, A. D. *et al.* The international workshop on osteoarthritis imaging knee MRI segmentation challenge: A multi-institute evaluation and analysis framework on a standardized dataset. *Radiol. Artif. Intell.* **3**, e200078 (2021).
33. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE International Conference on Computer Vision* 1026–1034 (2015).
34. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th International Conference on Machine Learning, ICML 2019* 10691–10700 (2019).
35. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
36. Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2014).
37. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* (2014).
38. Klinger, E. V. *et al.* Accuracy of race, ethnicity, and language preference in an electronic health record. *J. Gen. Intern. Med.* **30**, 719–723 (2015).
39. Glasheen, W. P. *et al.* Charlson comorbidity index: ICD-9 update and ICD-10 translation. *Am. Health Drug Benef.* **12**, 188 (2019).
40. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).
41. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. Preprint at https://arxiv.org/abs/1312.6034 (2013).
42. Abd Rahni, A. A., Mohamed Fuzaie, M. F. & Al Irr, O. I. Automated bed detection and removal from abdominal CT images for automatic segmentation applications. In *2018 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2018—Proceedings* 677–679 (2019).
43. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
44. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
45. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
46. 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).
47. 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 Dig. Health* **2**, e192–e200 (2020).
48. Narayan, A. K., Lopez, D. B., Kambadakone, A. R. & Gervais, D. A. Nationwide, longitudinal trends in CT colonography utilization: Cross-sectional survey results from the 2010 and 2015 National Health Interview Survey. *J. Am. Coll. Radiol.* **16**, 1052–1057 (2019).
49. Magudia, K. *et al.* Utility of normalized body composition areas, derived from outpatient abdominal CT using a fully automated deep learning method, for predicting subsequent cardiovascular events. *Am. J. Roentgenol.* **220**, 236–244 (2023).
50. Muntner, P. *et al.* Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *J. Am. Med. Assoc.* **311**, 1406–1415 (2014).
51. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. In *ICML'06: Proceedings of the 23rd International Conference on Machine Learning* 233–240 (2006).
52. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36. https://doi.org/10.1038/s41591-018-0307-0 (2019).
53. Weston, A. D. *et al.* Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology* **290**, 669–679 (2019).
54. Magudia, K. *et al.* Population-scale CT-based body composition analysis of a large outpatient population using deep learning to derive age-, sex-, and race-specific reference curves. *Radiology.* https://doi.org/10.1148/radiol.2020201640 (2020).
55. Forrest, I. S. *et al.* Machine learning-based marker for coronary artery disease: Derivation and validation in two longitudinal cohorts. *Lancet* **401**, 215–225 (2023).
56. Erickson, S. R., Cole, E., Kline-Rogers, E. & Eagle, K. A. The addition of the Charlson comorbidity index to the GRACE Risk prediction index improves prediction of outcomes in acute coronary syndrome. *Popul. Health Manag.* **17**, 54–59 (2014).
57. Núñez, J. E. *et al.* Prognostic value of Charlson comorbidity index at 30 days and 1 year after acute myocardial infarction. *Rev. Esp. Cardiol. (English Ed.)* **57**, 842–849 (2004).
58. Karabağ, T. *et al.* The relationship of Charlson comorbidity index with stent restenosis and extent of coronary artery disease. *Interv. Med. Appl. Sci.* **10**, 70 (2018).
59. Weber, T. *et al.* Hypertension and coronary artery disease: Epidemiology, physiology, effects of treatment, and recommendations: A joint scientific statement from the Austrian Society of Cardiology and the Austrian Society of Hypertension. *Wiener Klinische Wochensch.* **128**, 467–479 (2016).
60. Afsar, B., Turkmen, K., Covic, A. & Kanbay, M. An update on coronary artery disease and chronic kidney disease. *Int. J. Nephrol.* **2014**, 7424. https://doi.org/10.1155/2014/767424 (2014).
61. Pesapane, F., Codari, M. & Sardanelli, F. Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* **2**, 6. https://doi.org/10.1186/s41747-018-0061-6 (2018).
62. Meskó, B. & Görög, M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Dig. Med.* **3**, 1–8. https://doi.org/10.1038/s41746-020-00333-z (2020).
63. Yadlowsky, S. *et al.* Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann. Intern. Med.* **169**, 20 (2018).
64. Navar-Boggan, A. M., Peterson, E. D., D'Agostino, R. B., Pencina, M. J. & Sniderman, A. D. Using age- and sex-specific risk thresholds to guide statin therapy: One size may not fit all. *J. Am. Coll. Cardiol.* **65**, 1633–1639 (2015).

65. Dalton, J. E. *et al.* Accuracy of cardiovascular risk prediction varies by neighborhood socioeconomic position: A retrospective cohort study. *Ann. Intern. Med.* **167**, 456 (2017).
66. Rosenblit, P. D. Extreme atherosclerotic cardiovascular disease (ASCVD) risk recognition. *Curr. Diabetes Rep.* **19**, 1–20 (2019).
67. Lemieux, I. *et al.* Total cholesterol/HDL cholesterol ratio vs LDL cholesterol/HDL cholesterol ratio as indices of ischemic heart disease risk in men: The Quebec cardiovascular study. *Arch. Intern. Med.* **161**, 2685–2692 (2001).
68. Cantin, B. *et al.* Is lipoprotein(a) an independent risk factor for ischemic heart disease in men? The Quebec cardiovascular study. *J. Am. Coll. Cardiol.* **31**, 519–525 (1998).
69. Mendall, M. A. *et al.* C-reactive protein: Relation to total mortality, cardiovascular mortality and cardiovascular risk factors in men. *Eur. Heart J.* **21**, 1584–1590 (2000).
70. Orringer, C. E. *et al.* The National Lipid Association scientific statement on coronary artery calcium scoring to guide preventive strategies for ASCVD risk reduction. *J. Clin. Lipidol.* **15**, 33–60 (2021).

Author contributions

J.M.Z.C. performed all experiments and wrote the manuscript. A.S.C. helped with experiments, writing the manuscript, creating figures, and manuscript editing. A.D. developed the segmentation model. A.W. performed segmentations and provided manuscript edits. R.B., D.M., F.R., A.S., B.J., R.C., G.K. and D.R. provided manuscript edits. B.N.P. designed the concept and study, oversaw the experiments, and edited the manuscript.

Funding

The authors make the following disclosures. This research used services provided by STARR, "STAnford medicine Research data Repository," a clinical data warehouse containing data from Stanford Health Care (SHC), the Stanford Children's Hospital (SCH), the University Healthcare Alliance (UHA) and Packard Children's Health Alliance (PCHA) clinics and other auxiliary data from radiology. STARR is developed and operated by Stanford Medicine Research IT team and is made possible by Stanford School of Medicine Research Office. JMZC is supported by a graduate fellowship award from Knight-Hennessy Scholars at Stanford University, and receives research support from GE Healthcare not related to this work. ASC has provided consulting services to Subtle Medical, Chondrometrics GmbH, Image Analysis Group, Edge Analytics, ICM Co., and Culvert Engineering; and is a shareholder of Subtle Medical, LVIS Corporation, and Brain Key; and is on the advisory board for Chondrometrics GmbH and Brain Key; and receives research support from GE Healthcare and Philips not related to this work. ATS receives funding from the National Heart, Blood, and Lung Institute (K23 HL151672-01). FR was funded by a career development award from the National Heart, Lung, and Blood Institute (K01 HL 144607) and the American Heart Association/Robert Wood Johnson Harold Amos Medical Faculty Development Program. RDB, DJM, FR, ASC and BNP receive funding from the National Institutes of Health (R01 HL167974).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-47895-y.

Correspondence and requests for materials should be addressed to B.N.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.