Appendix

eMethods

Predictors selection

Ten predictors were finally selected from a deliberately designed data-driven strategy from a comprehensive space covering 645 candidate variables. The strategy consists of two main steps, variable importance ranking, and sequential forward selection.

Variable importance ranking was calculated using a built-in function within the LGBM algorithm [1]. As the LGBM is a tree-based model that contains a bunch of decision tree models, the variable importance can be measured by the number of that variable taken as split nodes, which is known as the model's "cover". The more frequently a variable is used during the tree constructions, the higher its relative importance to the model. The importance can be calculated explicitly for each feature in the whole dataset, allowing them to be ranked and compared to each other. Further, to diminish the bias that might be resulted from using a single set of hyperparameters, we arbitrarily trained 100 models under different parameter spaces and chose the top 5% (5 out of 100) of them based on the AUC. The final variable importance score was obtained by averaging those from the five best-performed models. After ranking the importance score of all candidate variables, we arbitrarily selected the top 50 ones.

Although the ensembled tree-based LGBM is tolerable to multicollinearity issues as it works by randomly selecting either of the highly correlated variables with no emphasis, it still may face the problem of low interpretation on final included predictors that several of them make similar or repeated contributions to model predictions. To alleviate the issue of multicollinearity, we calculated Spearman rank-order correlations [2] (eFigure 1a) to the 50 pre-selected variables and then converted the correlation matrix to a distance matrix defined by

$$dist_{matrix} = 1 - \frac{|\, corr_{matrix} + transpose(corr_{matrix})|}{2}$$

We then performed hierarchical clustering of the distance matrix to group variables based on Ward's linkage [3] (eFigure 1b). We used 0.75 as a threshold to cut the dendrogram and chose the best representative variable within each cluster. Among the 50 pre-selected variables, 28 of them surpass the hierarchical clustering and forward to the next procedure.

To further select optimal variables for ML model development, we employed a sequential forward selection strategy. We repeated the variable importance ranking procedure on the 28 variables, and consecutively develop classifiers by sequentially adding variables one by each iteration. The selection scheme can be delineated by the

line chart in Figure 2a that the model's performance climbed steeply when involved in the first couple of variables and gradually went to a plateau when additional ones joined in. Finally, we chose the top 10 variables as the final predictor for further model development.

CVD Risk Model development

The CVD risk model development consists of two steps: ML model development and risk calibration.

Our study adopted multiple popular ML algorithms and the LGBM (light gradient boosting machine) achieved the best performance according to eTable 6. LGBM is an example of ensemble learning methods that are constructed based on numerous underlying base learners, e.g., decision trees, to capture complex and non-linear patterns. The algorithm works by starting from a weak classifier (decision tree model) and consecutively building each new tree to correct the errors from the pre-trained ones. Such structure sequentially grows with the most promising branches and leaves, and finally produces a strong overall predictive model. In the prediction process, LGBM aggregates the probabilities derived from each individual decision tree to output an ensembled probability of a participant being classified into either incident of cardiovascular disease or staying healthy in the next 10 years.

The output of an ML model merely represents the probability of discriminating whether a participant can develop CVD, and a further step of calibration is required to map the raw probability to the processed probability (calibrated risks) in a cohort under specific prevalence. Thus, by using the output probabilities of ML models, we adopted isotonic regression [4, 5] to regress the output probabilities of ML models to the actual observed risk. As such, we aimed to assess the level of agreement between calibrated risks and observed proportions of CVD events. We drew the calibration plots based on decile subgroups. To be specific, the risks of all participants were sorted and partitioned into 10% quantile subgroups, the mean risk and observed proportions of events were then calculated within each subgroup. Under such a scheme, the risks were distributed in a monotoned increasing trend, and the observed proportions were expected to distribute in the same manner. We further calculated the Brier score [6] for the assessment of output risks versus proportions of actual observed events.

Leave-one-center-out cross-validation

Participants collected in the UK-Biobank cohort were recruited from 22 assessment centers across the UK. We split the dataset into 22 subsets based on the assessment centers (Field ID). Notably, the number of participants registered at centers in Stockport (n=3,554), Swansea (n=2,121), and Wrexham (n=620) were less than 1% of the whole study population (n=473,610); thus, we merged these participants in case of insufficient amount of incident target events. Thus, the cohort was partitioned into 20 sub-folds for model development and validation.

Each time 19 folds of data were used as a training set and the rest fold as a validation set; we repeated this process 20 times by shifting the folds of data as training and validation sets. Specifically, hyperparameters optimization and isotonic regression (risk calibration processor) were performed under inner-loop five-fold cross-validation in the training sets, and the validation sets were merely used for model evaluations. Reported results were calculated across the folds by using the averaged statistics.

Hyperparameters optimization of ML models

The performance of ML models relies heavily on the choices of the hyperparameter space, and we optimize the selection process based on maximizing the AUC based on inner-looped cross-validation within training sets. We adopted a grid search strategy by exploring all possible combinations within a pre-defined hyperparameter space, which is listed in eTable 3. The finally used hyperparameters to develop the UKCRP (LGBM model) were "n_estimators": 500; "num_leaves": 10; "max_depth": 15; "subsample": 0.7; "learning_rate": 0.01; "colsample_bytree": 0.7. For further supporting information on these parameters, please refer to the webpage of LGBM's documentation (https://github.com/microsoft/LightGBM).

Data Pre-processing

The LGBM algorithm supports missing values by default. In a tree-based model, split directions for missing values can be automatically learned during training. As the missingness is not tolerable to the rest ML classifiers and existing CVD risk prediction scales, imputation was performed. We conducted simple imputations based on participants' sex (mean for continuous variables and mode for discrete variables) for the variables with missingness less than 5% and multiple imputations for variables with missingness over 5%. The multiple imputations were conducted using a package of Miss forest [7, 8] under Python (v3.9).

ML algorithms of artificial neural networks (ANN), K-nearest-neighbors (KNN) and support vector machine (SVM) are sensitive to the scale of input variables; thus, necessary pre-processing steps of standardization (continuous variables) and one-hot-encoding (discrete variables) were conducted before the training procedures. No pre-processing is required by classifiers of tree-based algorithms, e.g., random forest, XGBoost and LGBM.

Polygenetic risk score (PRS) generation

Imputation data were available for all 487,409 participants in the UK Biobank cohort. Before calculating PRS, all samples and genotypes underwent stringent quality control. Specifically, SNPs were excluded if they had missing rate > 5%, minimum minor allele frequency (MAF) < 0.1%, or Hardy–Weinberg equilibrium test $P < 1\times10\text{-}50$. To minimize the variability due to population structure, we restricted our analyses to unrelated individuals based on the following three criteria: (1) not marked as outliers for heterozygosity and missing rates, (2) do not show

putative sex chromosome aneuploidy, (3) have at most ten putative third-degree relatives. After the quality control procedures, we obtained a total of 16,421,481 SNPs and 406,761 participants.

We calculated the PRS with the summary statistics from a meta-analysis of GWAS [9] for any stroke, comprising ischemic stroke, intracerebral hemorrhage, and stroke of unknown or undetermined type. This meta-analysis provided a total of 446,696 European participants (40,585 cases and 406,111 controls). The summary statistics we used included 8,255,860 SNPs that were present in the GWAS data. PRS were calculated using the PRSice software (www.PRSice.info). P-value-informed clumping with a cutoff of r2 = 0.1 in a 250-kb window was used in the analysis. A p-value threshold (PT) was used for the selection of the SNPs. Since the optimal P value threshold is unknown a priori, high-resolution PRSs are calculated over 100 p-value thresholds (PT, ranging from 0 to 0.5 with increments of 0.005).

eReference

1. Ke, G., et al., *Lightgbm: A highly efficient gradient boosting decision tree.* Advances in neural information processing systems, 2017. **30**.
2. Zwillinger, D. and S. Kokoska, *CRC standard probability and statistics tables and formulae*. 1999: Crc Press.
3. Rokach, L. and O. Maimon, *Clustering methods*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 321-352.
4. Chakravarti, N., *Isotonic median regression: a linear programming approach.* Mathematics of operations research, 1989. **14**(2): p. 303-308.
5. De Leeuw, J., *Correctness of Kruskal's algorithms for monotone regression with ties.* Psychometrika, 1977. **42**(1): p. 141-144.
6. Brier, G.W., *Verification of forecasts expressed in terms of probability.* Monthly weather review, 1950. **78**(1): p. 1-3.
7. Van Buuren, S. and K. Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in R.* Journal of statistical software, 2011. **45**: p. 1-67.
8. Little, R.J. and D.B. Rubin, *Statistical analysis with missing data*. Vol. 793. 2019: John Wiley & Sons.
9. Malik, R., et al., *Publisher correction: multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes.* Nature genetics, 2019. **51**(7): p. 1192-1193.

eTable 1: Candidate variables from UK-Biobank

Category	UK-Biobank Field IDs & Self-generated variables
Biofluid assays (n = 70)	30730-0.0, 30740-0.0, 30790-0.0, 30890-0.0, 30610-0.0, 30830-0.0, 30680-0.0, 30860-0.0, 30620-0.0, 30600-0.0, 30760-0.0, 30770-0.0, 30840-0.0, 30630-0.0, 30700-0.0, 30660-0.0, 30710-0.0, 30720-0.0, 30750-0.0, 30870-0.0, 30640-0.0, 30670-0.0, 30880-0.0, 30650-0.0, 30810-0.0, 30690-0.0, 30780-0.0, 30850-0.0, 30190-0.0, 30210-0.0, 30290-0.0, 30030-0.0, 30010-0.0, 30110-0.0, 30170-0.0, 30260-0.0, 30230-0.0, 30220-0.0, 30060-0.0, 30180-0.0, 30300-0.0, 30200-0.0, 30070-0.0, 30150-0.0, 30250-0.0, 30160-0.0, 30270-0.0, 30100-0.0, 30050-0.0, 30020-0.0, 30240-0.0, 30140-0.0, 30280-0.0, 30040-0.0, 30120-0.0, 30090-0.0, 30130-0.0, 30080-0.0, 30000-0.0, 30354-0.0, 30314-0.0, 30520-0.0, 30510-0.0, 30500-0.0, 30530-0.0, 30503-0.0, 30523-0.0, 30533-0.0, 30513-0.0, CHOL_RATIO
Cognitive function (n = 71)	20016-0.0, 20018-0.0, 20023-0.0, 20128-0.0, 396-0.1, 396-0.2, 397-0.1, 397-0.2, 398-0.1, 398-0.2, 399-0.1, 399-0.2, 400-0.1, 400-0.2, 401-0.0, 401-0.1, 401-0.10, 401-0.11, 401-0.2, 401-0.3, 401-0.4, 401-0.5, 401-0.6, 401-0.7, 401-0.8, 401-0.9, 402-0.0, 402-0.1, 402-0.10, 402-0.11, 402-0.2, 402-0.3, 402-0.4, 402-0.5, 402-0.6, 402-0.7, 402-0.8, 402-0.9, 403-0.0, 403-0.1, 403-0.10, 403-0.11, 403-0.2, 403-0.3, 403-0.4, 403-0.5, 403-0.6, 403-0.7, 403-0.8, 403-0.9, 404-0.0, 404-0.1, 404-0.10, 404-0.11, 404-0.2, 404-0.3, 404-0.4, 404-0.5, 404-0.7, 4287-0.0, 4288-0.0, 4290-0.0, 4291-0.0, 4292-0.0, 4293-0.0, 4294-0.0, 4924-0.0, 4935-0.0, 4946-0.0, 4957-0.0, 4968-0.0
Early life factors (n = 10)	120-0.0, 1647-0.0, 1677-0.0, 1687-0.0, 1697-0.0, 1707-0.0, 1767-0.0, 1777-0.0, 1787-0.0, 20022-0.0
Family history (n = 28)	1797-0.0, 1807-0.0, 1835-0.0, 1873-0.0, 1883-0.0, 20107-0.0, 20107-0.1, 20110-0.0, 20110-0.1, 20111-0.0, 20111-0.1, 3526-0.0, 4501-0.0, sibling_diab, sibling_hbp, sibling_str, sibling_hd, sibling_cvd, parent_diab, parent_hbp, parent_str, parent_hd, parent_cvd, family_diab, family_hbp, family_str, family_hd, family_cvd
Health and medical history (n = 46)	134-0.0, 135-0.0, 136-0.0, 20009-0.0, 20011-0.0, 2188-0.0, 2207-0.0, 2217-0.0, 2227-0.0, 2247-0.0, 2257-0.0, 2296-0.0, 2306-0.0, 2316-0.0, 2335-0.0, 2345-0.0, 2355-0.0, 2443-0.0, 2453-0.0, 2463-0.0, 2473-0.0, 2492-0.0, 2966-0.0, 3393-0.0, 3571-0.0, 4717-0.0, 4728-0.0, 4792-0.0, 4803-0.0, 4825-0.0, 4836-0.0, 6148-0.0, 6149-0.0, 6152-0.0, 6155-0.0, 6159-0.0, 6159-0.1, 6179-0.0, 87-0.0, HYPT, AF, HeartAttack, HighBP, ANGINA, ANG_HA, ChestPain
Lifestyle and environment (n = 143)	1011-0.0, 1021-0.0, 1050-0.0, 1060-0.0, 1070-0.0, 1080-0.0, 1090-0.0, 1100-0.0, 1110-0.0, 1120-0.0, 1130-0.0, 1140-0.0, 1150-0.0, 1160-0.0, 1170-0.0, 1180-0.0, 1190-0.0, 1200-0.0, 1210-0.0, 1220-0.0, 1259-0.0, 1289-0.0, 1299-0.0, 1309-0.0, 1319-0.0, 1329-0.0, 1339-0.0, 1349-0.0, 1359-0.0, 1369-0.0, 1379-0.0, 1389-0.0, 1408-0.0, 1418-0.0, 1428-0.0, 1438-0.0, 1448-0.0, 1458-0.0, 1468-0.0, 1478-0.0, 1488-0.0, 1498-0.0, 1508-0.0, 1518-0.0, 1528-0.0, 1538-0.0, 1548-0.0, 1558-0.0, 1568-0.0, 1578-0.0, 1588-0.0, 1598-0.0, 1608-0.0, 1618-0.0, 1628-0.0, 1717-0.0, 1727-0.0, 1737-0.0, 1747-0.0, 1757-0.0, 20117-0.0, 20160-0.0, 20161-0.0, 20162-0.0, 2129-0.0, 2139-0.0, 2149-0.0, 2159-0.0, 22032-0.0, 22033-0.0, 22034-0.0, 22035-0.0, 22036-0.0, 22037-0.0, 22038-0.0, 22039-0.0, 22040-0.0, 2237-0.0, 2267-0.0, 2277-0.0, 24003-0.0, 24004-0.0, 24005-0.0, 24006-0.0, 24007-0.0, 24008-0.0, 24009-0.0, 24010-0.0, 24011-0.0, 24012-0.0, 24013-0.0, 24014-0.0, 24015-0.0, 24016-0.0, 24017-0.0, 24018-0.0, 24019-0.0, 24020-0.0, 24021-0.0, 24022-0.0, 24023-0.0, 24024-0.0, 24500-0.0, 24501-0.0, 24502-0.0, 24503-0.0, 24504-0.0, 24505-0.0, 24506-0.0, 24507-0.0, 24508-0.0, 2624-0.0, 2634-0.0, 2654-0.0, 2664-0.0, 2867-0.0, 2877-0.0, 2897-0.0, 2907-0.0, 2926-0.0, 2936-0.0, 3637-0.0, 3647-0.0, 6144-0.0, 6157-0.0, 6162-0.0, 6162-0.1, 6164-0.0, 6164-0.1, 6164-0.2, 864-0.0, 874-0.0, 884-0.0, 894-0.0, 904-0.0, 914-0.0, 924-0.0, 943-0.0, 971-0.0, 981-0.0, SMK_EXP, SMK_STAT, SMK_QT_YRS
Medications (n = 9)	137-0.0, CL_MED, BP_MED, CL_BP_MED, IN_MED, PAIN_MED, ASP_MED, LBU_MED, PAR_MED
Physical measures (n = 197)	23114-0.0, 23130-0.0, 23124-0.0, 23117-0.0, 23104-0.0, 23115-0.0, 23126-0.0, 23125-0.0, 23105-0.0, 23099-0.0, 23123-0.0, 23119-0.0, 23122-0.0, 23121-0.0, 23118-0.0, 23128-0.0, 23109-0.0, 23098-0.0, 23113-0.0, 23106-0.0, 23120-0.0, 23107-0.0, 23100-0.0, 23112-0.0, 23116-0.0, 23110-0.0, 23129-0.0, 23101-0.0, 23102-0.0, 23108-0.0, 23111-0.0, 23127-0.0, 48-0.0, 3077-0.0, 50-0.0, 49-0.0, 20015-0.0, 21-0.0, 21001-0.0, 51-0.0, 21002-0.0, 102-0.0, 102-0.1, 4079-0.0, 4080-0.1, 4093-0.0, 4119-0.0, 4101-0.0, 4095-0.0, 3082-0.0, 3148-0.0, 3147-0.0, 3144-0.0, 4125-0.0, 4124-0.0, 78-0.0, 4100-0.0, 4104-0.0, 3081-0.0, 4120-0.0, 4096-0.0, 4105-0.0, 4123-0.0, 4092-0.0, 19-0.0, 3143-0.0, 4106-0.0, 46-0.0, 47-0.0, 4233-0.1, 4244-0.11, 4243-0.13, 4230-0.14, 4241-0.10, 4230-0.8, 4243-0.14, 4230-0.10, 4244-0.4, 4268-0.0, 4244-0.10, 4241-0.9, 4233-0.2, 4241-0.15, 4241-0.12, 4241-0.8, 4233-0.12, 4230-0.3, 4241-0.1, 4232-0.3, 4243-0.4, 4232-0.12, 4230-0.7, 4233-0.11, 4275-0.0, 4276-0.0, 4232-0.13, 4244-0.1, 4230-0.1, 4232-0.4, 4232-0.11, 4269-0.0, 4241-0.11, 4243-0.10, 4241-0.3, 4230-0.2, 4241-0.2, 4233-0.15, 4243-0.12, 4244-0.2, 4244-0.8, 4244-0.6, 4232-0.5, 4232-0.10, 4277-0.0, 4232-0.1, 4244-0.14, 4243-0.8, 4232-0.6, 20021-0.0, 4244-0.3, 4233-0.9, 4230-0.12, 4244-0.9, 4232-0.14, 4232-0.15, 4241-0.6, 4233-0.8, 4241-0.4, 4230-0.13, 4241-0.5, 4243-0.7, 4241-0.13, 4230-0.15, 4243-0.1, 4244-0.12, 4230-0.6, 4232-0.8, 4243-0.11, 4233-0.14, 4849-0.0, 4244-0.13, 4230-0.11, 4244-0.5, 4233-0.10, 4233-0.13, 4241-0.7, 4244-0.15, 4230-0.9, 4233-0.5, 4243-0.15, 4233-0.3, 4244-0.7, 4232-0.9, 4243-0.9, 4243-0.3, 4230-0.5, 4230-0.4, 4233-0.4, 4243-0.2, 4243-0.6, 4270-0.0, 4232-0.7, 4233-0.7, 4241-0.14, 4243-0.5, 4233-0.6, 20019-0.0, 4232-0.2, 3063-0.2, 3059-0.2, 20257-0.0, 3065-0.2, 3062-0.1, 3063-0.0, 20154-0.0, 3065-0.1, 20152-0.0, 3090-0.0, 20256-0.0, 3063-0.1, 23-0.0,

	20150-0.0, 20153-0.0, 3064-0.2, 3088-0.0, 3064-0.0, 3062-0.2, 3089-0.0, 20258-0.0, 3062-0.0, 20255-0.0, 3137-0.0, 3065-0.0, 3059-0.0, 20151-0.0, 3059-0.1, 3064-0.1
Psychosocial factors (n = 34)	1940-0.0, 4653-0.0, 2040-0.0, 4559-0.0, 1970-0.0, 2090-0.0, 1980-0.0, 4598-0.0, 4581-0.0, 2020-0.0, 4570-0.0, 2050-0.0, 20127-0.0, 1950-0.0, 2070-0.0, 2010-0.0, 2030-0.0, 1960-0.0, 4537-0.0, 6145-0.0, 1930-0.0, 2000-0.0, 4642-0.0, 4631-0.0, 1920-0.0, 4548-0.0, 2080-0.0, 4526-0.0, 1990-0.0, 2100-0.0, 2060-0.0, 1031-0.0, 6160-0.0, 2110-0.0
Socio-demographics (n = 37)	52-0.0, 189-0.0, 31-0.0, 21022-0.0, 26410-0.0, 26411-0.0, 26415-0.0, 26417-0.0, 26416-0.0, 26414-0.0, 26413-0.0, 26412-0.0, 6138-0.0, 6138-0.1, 845-0.0, 826-0.0, 767-0.0, 816-0.0, 796-0.0, 6142-0.0, 806-0.0, 777-0.0, 757-0.0, 6143-0.0, 21000-0.0, 670-0.0, 6140-0.0, 680-0.0, 6139-0.0, 709-0.0, 738-0.0, 728-0.0, 699-0.0, 6141-0.0, 6139-0.1, 6146-0.0, 4674-0.0

Please refer to the webpage of UK-Biobank for detailed information (https://www.ukbiobank.ac.uk) on each variable. Variables except Field ID number were features that were not directly available from the database and further manually self-generated based on a combination of two or more ones. Detailed notations were given in eTable 2.

eTable 2: Notation table of self-generated variables

Category	Variables	Notations	Derived Field IDs
Biofluid assays	CHOL_RATIO	Ratio of total-cholesterol/ HDL-cholesterol	30690-0.0, 30760-0.0
Family history	sibling_diab	Diabetes of sibling	{20111-0.0, 20111-0.1, …, 20111-0.11}
	sibling_hbp	High blood pressure of sibling	{20111-0.0, 20111-0.1, …, 20111-0.11}
	sibling_str	Stroke of sibling	{20111-0.0, 20111-0.1, …, 20111-0.11}
	sibling_hd	Heart disease of sibling	{20111-0.0, 20111-0.1, …, 20111-0.11}
	sibling_cvd	Cardiovascular disease of sibling	{20111-0.0, 20111-0.1, …, 20111-0.11}
	parent_diab	Diabetes of parents	{20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	parent_hbp	High blood pressure of parents	{20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	parent_str	Stroke of parents	{20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	parent_hd	Heart disease of parents	{20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	parent_cvd	Cardiovascular disease of parents	{20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	family_diab	Diabetes of family members	{20111-0.0, 20111-0.1, …, 20111-0.11}, {20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	family_hbp	High blood pressure of family members	{20111-0.0, 20111-0.1, …, 20111-0.11}, {20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	family_str	Stroke of family members	20111-0.0, 20111-0.1, …, 20111-0.11}, {20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	family_hd	Heart disease of family members	{20111-0.0, 20111-0.1, …, 20111-0.11}, {20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
	family_cvd	Cardiovascular disease of family members	{20111-0.0, 20111-0.1, …, 20111-0.11}, {20107-0.0, 20107-0.1, …, 20107-0.9}, {20110-0.0, 20110-0.1, …, 20110-0.10}
Health and medical history	HYPT	Previous essential hypertension	131286-0.0, 131287-0.0, 53-0.0
	AF	Previous atrial fibrillation	131350-0.0, 131351-0.0, 53-0.0
	HeartAttack	Previous heart attack	6150-0.0, 6150-0.1, 6150-0.2, 6150-0.3
	HighBP	Previous high blood pressure	6150-0.0, 6150-0.1, 6150-0.2, 6150-0.3
	ANGINA	Previous anginal	131296-0.0, 131297-0.0, 53-0.0, 6150-0.0, 6150-0.1, 6150-0.2, 6150-0.3
	ANG_HA	Angina or heart attack	131296-0.0, 131297-0.0, 53-0.0, 6150-0.0, 6150-0.1, 6150-0.2, 6150-0.3
	ChestPain	Chest pain or discomfort	2335-0.0, 3606-0.0, 3616-0.0, 3751-0.0
Lifestyle and environment	SMK_EXP	Smoking exposure (hr/ week)	1269-0.0, 1279-0.0
	SMK_STAT	Smoking status (five leveled)	1239-0.0, 1249-0.0
	SMK_QT_YRS	Years after quit smoking (up-to-baseline)	21022-0.0, 2897-0.0
Medications	CL_MED	Cholesterol medication	6153-0.0, 6153-0.1, 6153-0.2, 6153-0.3, 6177-0.0, 6177-0.1, 6177-0.2
	BP_MED	Blood pressure medication	6153-0.0, 6153-0.1, 6153-0.2, 6153-0.3, 6177-0.0, 6177-0.1, 6177-0.2
	CL_BP_MED	Cholesterol & blood pressure medication	6153-0.0, 6153-0.1, 6153-0.2, 6153-0.3, 6177-0.0, 6177-0.1, 6177-0.2
	IN_MED	Insulin medication	6153-0.0, 6153-0.1, 6153-0.2, 6153-0.3, 6177-0.0, 6177-0.1, 6177-0.2
	PAIN_MED	Pain relief medication	{10004-0.0, 10004-0.1, …, 10004-0.4}
	ASP_MED	Aspirin medication	{10004-0.0, 10004-0.1, …, 10004-0.4}
	LBU_MED	Ibuprofen medication	{10004-0.0, 10004-0.1, …, 10004-0.4}
	PAR_MED	Paracetamol medication	{10004-0.0, 10004-0.1, …, 10004-0.4}

Self-generated variables cannot be directly accessed from the UKB dataset and were derived from combinations of two or more available ones. Decisions to create these variables were based on empirical knowledge. Field IDs shown in brace ({…}) are from one Field with multiple arrays.

eTable 3: Hyperparameter space explored for different machine learning classifiers

ML classifiers	Hyperparameters	Range	Step	Final choice
KNN	n_neighbors	{10, 100}	10	90
	weights	{'uniform', 'distance'}		'distance'
	algorithm	{'auto', 'ball_tree', 'kd_tree', 'brute'}		'kd_tree'
Logistic regression	solver	{'newton-cg', 'liblinear'}	/	'newton-cg'
	penalty	{none, l1, l2}	/	l2
	C	{0.005, 0.01, 0.05, 0.1, 0.5, 1}	/	1
SVM	kernel	{'rbf', 'sigmoid'}	/	'rbf'
	C	{0.0001, …, 10000}	*10	1000
	gamma	{'scale', 'auto', 0.0001, 0.001, 0.01, 0.1}	/	0.001
Random forest	n_estimators	{100, …, 1000}	100	500
	criterion	'gini', 'entropy'	/	entropy
	max_depth	{3, …, 15}	2	7
	min_samples_leaf	{3, …, 15}	2	3
	min_samples_split	{3, …, 15}	2	7
	max_features	{'auto', 'sqrt', 'log2'}	/	'log2'
LGBM	n_estimators	{100, …, 1000}	100	500
	max_depth	{3, …, 30}	3	15
	subsample	{0.7, …, 1}	0.05	0.7
	colsample_bytree	{0.7, …, 1}	0.05	1
	learning_rate	{1e-5, …, 1e-1}	*10	1e-2
	num_leaves	{10, …, 100}	10	10
XGBoost	n_estimators	{100, …, 1000}	100	500
	max_depth	{3, …, 15}	3	6
	min_child_weight	{3, …, 15}	3	3
	subsample	{0.7, …, 1}	0.05	0.9
	eta	{1e-5, …, 1e-1}	*10	1e-2
ANN	Learning rate	{1e-5, …, 1e-1}	*10	1e-3
	Number of layers	{1, …, 5}	1	3
	Layer size	{3, 5, 7, 10}	/	5
	Batch size	{128, …, 1024}	*2	256
	Epochs	{10, …, 100}	10	10
	Dropout	{0, …, 0.5}	0.05	0.3
	optimizer	{'Adam', 'Adamax', 'SGD', 'RMSprop'}	/	'Adam'

Abbreviations: ANN = Artificial Neural Network, KNN = K-nearest-neighbours, LGBM = Light Gradient Boosting Machine, SVM = Support Vector Machine, XGBoost = eXtreme Gradient Boosting Machine.

eTable 4: Study population and modelling algorithms of our study versus existing prediction scales

	UKCRP	QRISK3	SCORE2	AHA/ASCVD	FGCRS
Publish year		2017	2021	2013	2007
Derivational Population	473,611	7,889,803	677,684	24,626	8,491
target events (%)	31,466 (6.6%)	363,565 (4.61%)	30,121 (4.44%)	2,690 (10.9%)	1,174 (13.8%)
Observation time (years)	12.2 IQR [11.5-12.9]	4.4 IQR [1.6-10.8]	10.7 5th/95th percentile [5.0-18.6]	>12	> 12
Mean age (years) [range]	56.4 [37-73]	43.0 [25-84]	57 [40-69]	50.2 [40-79]	49 [30-74]
Sex (females)	264,308 (55.8%)	4,019,956 (51.0%)	376,949 (55.6%)	13,881 (56.4%)	4,522 (53.3%)
Model algorithm	LGBM	Cox regression	Cox regression	Cox regression	Cox regression
Number of predictors used	10	21	7	8	8
Predictors used	Age, sex, cholesterol or blood pressure treatment, cholesterol ratio (Total/HDL), systolic blood pressure (SBP), angina or heart attack, number of medications, cystatin C, chest pain, pack-year of smoking	Age, sex, Townsend score, ethnicity, smoking status, height, weight, systolic blood pressure (SBP), blood pressure treatment, cholesterol ratio (Total/HDL), diabetes status, angina or heart attack, chronic kidney disease, atrial fibrillation, migraines, rheumatoid arthritis, systemic lupus erythematosus (SLE), severe mental illness, atypical antipsychotic medication, steroid medication, erectile dysfunction	Age, sex, current smoker, systolic blood pressure (SBP), total cholesterol, HDL-cholesterol, risk regions	Age, sex, ethnicity, current smoker, systolic blood pressure (SBP), total cholesterol, HDL-cholesterol, diabetes mellitus, blood pressure treatment	Age, sex, current smoker, systolic blood pressure (SBP), total cholesterol, HDL-cholesterol, diabetes mellitus, blood pressure treatment

eTable 5: Summary statistics of 10-year incident myocardial infarction, ischemic stroke, and hemorrhagic stroke

Participants Characteristics	Healthy control (MI) (n=447,977)	Myocardial infarction (n=25,634)	Healthy control (IS) (n=468,000)	Ischemic stroke (n=56,11)	Healthy control (HS) (n=471,924)	Hemorrhagic stroke (n=1,687)
Age, year	57 [49-63]	62 [57-66]	57 [50-63]	63 [58-66]	57 [50-63]	62 [56-66]
Sex (female)	255258 (57.0%)	9050 (35.3%)	262006 (56.0%)	2302 (41.0%)	263389 (55.8%)	919 (54.1%)
Ethnicity (White)	421134 (94.0%)	23941 (93.4%)	439774 (94.0%)	5301 (94.5%)	443473 (94.0%)	1602 (94.3%)
Systolic blood pressure (mmHg)	134 [122-147]	141 [129-154]	134 [123-147]	143 [130-156]	134 [123-147]	141 [129-154]
Total cholesterol (mmol/L)	5.71 [5.00-6.46]	5.62 [4.78-6.48]	5.70 [4.99-6.46]	5.64 [4.83-6.44]	5.70 [4.99-6.46]	5.68 [4.93-6.44]
HDL-cholesterol (mmol/L)	1.42 [1.19-1.69]	1.27 [1.08-1.51]	1.41 [1.18-1.69]	1.32 [1.11-1.59]	1.41 [1.18-1.68]	1.44 [1.18-1.71]
Cholesterol-ratio (Total/HDL)	3.95 [3.36-4.71]	4.34 [3.62-5.13]	3.97 [3.3-4.73]	4.19 [3.52-4.93]	3.97 [3.37-4.73]	3.95 [3.31-4.72]
Cystatin C	0.88 [0.8-0.97]	0.95 [0.86-1.06]	0.88 [0.80-0.97]	0.96 [0.86-1.07]	0.88 [0.8-0.98]	0.91 [0.82-1.02]
Chest pain	20535 (4.6%)	3591 (10.8%)	23615 (5.0%)	511 (9.1%)	24008 (5.1%)	118 (6.9%)
Current smoker	45287 (10.1%)	3848 (15.0%)	48210 (10.3%)	925 (16.5%)	48878 (10.4%)	257 (15.1%)
Pack years of smoking	18.0 [9.3-30.0]	25.0 [13.8-39.5]	18.2 [9.5-30.8]	27.0 [14.5-41.3]	18.4 [9.5-31.0]	22.5 [11.5-37.5]
Cholesterol & blood pressure treatment						
either	69183 (15.4%)	6835 (26.7%)	74626 (15.9%)	1392 (24.8%)	75657 (16.0%)	361 (21.3%)
both	32241 (7.2)	5135 (20%)	36447 (7.8%)	929 (16.6%)	37169 (7.9%)	207 (12.2%)
Number of medications	2.0 [0.0-3.0]	3.0 [1.0-5.0]	2.0 [0.0-3.0]	3.0 [1.0-5.0]	2.0 [0.0-3.0]	2.0 [1.0-4.0]
Angina or heart attack	4993 (1.1%)	2489 (9.7%)	7272 (1.6%)	210 (3.7%)	7437 (1.6%)	45 (2.7%)
Diabetes	18563 (4.1%)	3038 (11.8%)	20984 (4.5%)	617 (11.0%)	21486 (4.6%)	115 (6.8%)
Hypertension	34156 (7.6%)	4832 (18.8%)	37992 (8.1%)	996 (17.8%)	38741 (8.2%)	247 (14.5%)

Data presented as median [IQR] for continuous variables and number (%) for discrete variables.

eTable 6: Model performance metrics for different machine learning classifiers on 10-year incident cardiovascular disease

	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
LGBM (UKCRP)	0.667±0.035	0.727±0.039	0.663±0.040	0.131±0.017	0.222±0.023	0.762±0.010
XGBoost	0.667 ±0.035	0.725 ±0.042	0.662 ±0.040	0.131 ±0.013	0.221 ±0.018	0.761 ±0.010
ANN	0.660 ±0.039	0.720 ±0.042	0.657 ±0.038	0.126 ±0.015	0.217 ±0.021	0.757 ±0.013
SVM	0.661 ±0.040	0.724 ±0.052	0.657 ±0.046	0.129 ±0.015	0.218 ±0.021	0.756 ±0.011
Random Forest	0.653 ±0.031	0.730 ±0.033	0.647 ±0.035	0.126 ±0.014	0.215 ±0.020	0.755 ±0.012
Logistic Regression	0.653 ±0.042	0.704 ±0.050	0.647 ±0.048	0.129 ±0.015	0.215 ±0.021	0.752 ±0.011
KNN	0.647 ±0.025	0.731 ±0.032	0.642 ±0.028	0.125 ±0.017	0.210 ±0.024	0.750 ±0.014

Binarization cut-off was determined based on the achievement of the largest Youden index (Youden index = sensitivity + specificity - 1).

Abbreviations: ANN = Artificial Neural Network, KNN = K-nearest-neighbours, LGBM = Light Gradient Boosting Machine, SVM = Support Vector Machine, XGBoost = eXtreme Gradient Boosting Machine.

eTable 7: Top-10 selected predictors by individually modeling on different outcome populations

Predictors	Target outcomes				Total count
	Cardiovascular disease	Myocardial infarction	Ischemic stroke	Hemorrhagic stroke	
Age	√	√	√	√	4
Systolic blood pressure (SBP)	√	√	√	√	4
Cystatin C	√	√	√	√	4
Sex	√	√	√		3
Pack years of smoking	√	√	√		3
Cholesterol ratio (total/HDL)	√	√			2
Cholesterol & blood pressure treatments	√	√			2
Previous angina or heart attack	√	√			2
Chest pain	√	√			2
Number of medications	√				1
Number of non-cancer illnesses		√			1
Hypertension (HBP)			√		1
Whole body fat-free mass			√		1
Microalbuminuria (MAU)			√		1
Albumin			√		1
Long-standing illness or disability			√		1
Mother's age at death				√	1
Forced expiratory volume (FEV) Z-score				√	1
Mean sphered cell volume (MSCV)				√	1
Cognitive reaction time				√	1
Limb fat percentage				√	1
Crime score				√	1
Forced expiratory volume (FEV)				√	1

Predictors listed in the left columns are the union set of top-10 selected predictors under individual modeling of the 10-year incident of cardiovascular disease (CVD) and its sub-diagnosis of myocardial infarction, ischemic stroke, and hemorrhagic stroke, respectively. The right column indicated how many times the predictor was selected. Age, systolic blood pressure, and cystatin C were chosen in all four models, followed by sex and smoking, which were chosen in three models. In general, selected predictors of myocardial infarction were largely consistent with those of CVD, which mainly resulted from its large proportion that over 80% of the CVD. Hemorrhagic stroke shares only three predictors to the CVD due to its small proportion in the target events and different pathogenesis to the other diseases.

eTable 8: Notation table of selected predictors

Selected predictors	Field ID	Category	Type	Notes
Age	21022-0.0	Socio-demographics	continuous	Age at baseline
Sex	31-0.0	Socio-demographics	discrete	0 = female; 1 = male
Cholesterol & blood pressure medication	/	Medications	discrete	0 = none; 1 = either; 2 = both. Derived based on the female specified variable "Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones" (Field ID 6153) and the male specified variable "Medication for cholesterol, blood pressure or diabetes" (Field ID 6177)
Cholesterol ratio (total/HDL)	/	Biofluid assays	continuous	Ratio of total-cholesterol (Field ID 30690)/ HDL-cholesterol (Field ID 30760)
Systolic blood pressure (SBP)	4080-0.0	Physical measures	continuous	Automated reading
Angina or heart attack	/	Health and medical history	discrete	0 = none; 1 = yes. Derived based on vascular/heart problems diagnosed by the doctor (Field ID 6150) and source of the report of I20 (Field ID 131296, 131297) any report before the baseline visit time (Field ID 53)
Number of medications	137-0.0	Medications	continuous	Number of medications (treatments) self-reported in the questionnaires
Cystatin C (mg/L)	30720-0.0	Biofluid assays	continuous	Measured by latex enhanced immunoturbidimetric analysis on a Siemens ADVIA 1800
Chest pain	/	Health and medical history	discrete	0 = none; 1 = Yes. Experienced any pain or discomfort in the chest (Field ID 2335, 3606, 3616, 3751)
Pack years of smoking	20161-0.0	Lifestyle and environment	continuous	Number of cigarettes per day / 20 * (Age stopped smoking - Age start smoking)

eTable 9: Odds ratio statistics of selected predictors

Selected predictors	Raw data				Normalized data			
	Odds ratio	95% Confidence Interval	Z-statistic	p-value	Odds ratio	95% Confidence Interval	Z-statistic	p-value
Age	1.06	[1.06-1.06]	58.74	<2.2e-16 ***	48.10	[42.26-54.75]	58.61	<2.2e-16 ***
Sex	1.90	[1.85-1.95]	47.15	<2.2e-16 ***	1.90	[1.85-1.95]	47.29	<2.2e-16 ***
Cholesterol & blood pressure medication								
either	1.31	[1.27-1.36]	16.82	<2.2e-16 ***	1.31	[1.27-1.35]	16.62	<2.2e-16 ***
both	1.45	[1.39-1.51]	17.58	<2.2e-16 ***	1.44	[1.38-1.50]	17.25	<2.2e-16 ***
Cholesterol ratio (total/HDL)	1.21	[1.20-1.22]	32.40	<2.2e-16 ***	3.19	[2.97-3.42]	32.24	<2.2e-16 ***
Systolic blood pressure (SBP)	1.01	[1.01-1.01]	26.72	<2.2e-16 ***	4.78	[4.26-5.36]	26.77	<2.2e-16 ***
Angina or heart attack	3.65	[3.45-3.87]	44.67	<2.2e-16 ***	3.64	[3.44-3.86]	44.48	<2.2e-16 ***
Number of medications	1.08	[1.08-1.09]	31.20	<2.2e-16 ***	1.76	[1.70-1.82]	31.39	<2.2e-16 ***
Cystatin C (mg/L)	1.98	[1.86-2.10]	21.28	<2.2e-16 ***	2.18	[2.03-2.35]	21.31	<2.2e-16 ***
Chest pain	1.89	[1.81-1.97]	29.35	<2.2e-16 ***	1.88	[1.80-1.96]	29.19	<2.2e-16 ***
Pack years of smoking	1.01	[1.01-1.01]	24.10	<2.2e-16 ***	1.35	[1.31-1.38]	24.20	<2.2e-16 ***

Odds ratios were calculated based on a multivariate logistic regression including all ten predictors. Two sets of odds ratios were reported based on inputs of data: non-normalized or normalized. Normalization were performed on continuous predictors by dividing their 99% quantile value to constrain their values between [0-1].

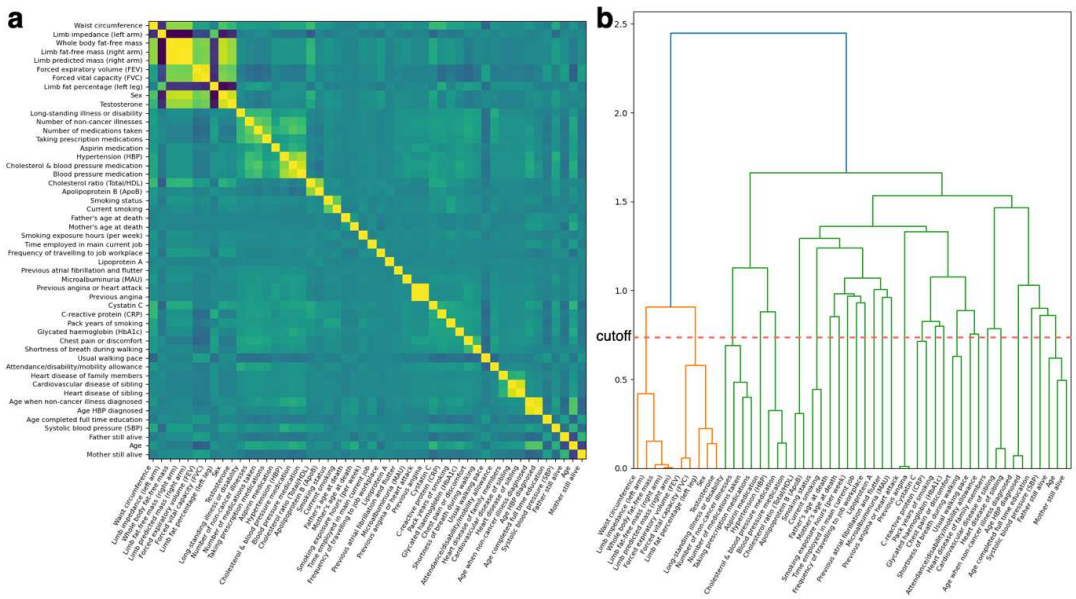
eTable 10: Model performance metrics for the prediction of 10-year incident cardiovascular disease and its sub-diagnostic groups

	Methods	Accuracy	Sensitivity	Specificity	Precision	F1-score	Brier score	AUC
Cardiovascular disease	UKCRP	0.667±0.035	0.727±0.039	0.663±0.040	0.131±0.017	0.222±0.023	0.057±0.006	0.762±0.010
	UKCRP+PRS	0.668±0.030	0.726±0.028	0.664±0.034	0.131±0.013	0.222±0.019	0.057±0.006	0.763±0.010
	QRISK3	0.647±0.027	0.724±0.026	0.642±0.030	0.124±0.013	0.211±0.019	0.058±0.006	0.744±0.011
	SCORE2	0.607±0.033	0.727±0.043	0.599±0.038	0.112±0.011	0.194±0.016	0.059±0.007	0.716±0.015
	AHA/ASCVD	0.601±0.050	0.708±0.044	0.593±0.057	0.109±0.010	0.188±0.015	0.059±0.007	0.701±0.014
	FGCRS	0.694±0.045	0.589±0.053	0.702±0.050	0.122±0.013	0.201±0.018	0.059±0.007	0.703±0.017
Myocardial infarction	UKCRP	0.671±0.038	0.739±0.044	0.668±0.042	0.111±0.015	0.192 ±0.022	0.047±0.006	0.774±0.011
	UKCRP+PRS	0.675±0.023	0.735±0.024	0.671±0.026	0.111±0.012	0.193±0.018	0.047±0.006	0.774±0.010
	Per Diagnosis	0.668±0.008	0.744±0.050	0.664±0.042	0.111±0.014	0.192±0.020	0.046±0.006	0.777±0.011
	QRISK3	0.644±0.026	0.736±0.027	0.639±0.029	0.102±0.012	0.179±0.018	0.048±0.006	0.750±0.013
	SCORE2	0.594±0.027	0.745±0.042	0.586±0.042	0.091±0.010	0.162±0.016	0.049±0.006	0.719±0.018
	AHA/ASCVD	0.606±0.046	0.701±0.049	0.601±0.051	0.089±0.010	0.158±0.015	0.049±0.006	0.702±0.016
	FGCRS	0.699±0.020	0.606±0.035	0.704±0.021	0.102±0.011	0.175±0.015	0.048±0.006	0.713±0.021
Ischemic stroke	UKCRP	0.644±0.062	0.710±0.061	0.643±0.0063	0.024±0.005	0.046±0.010	0.012±0.002	0.730±0.020
	UKCRP+PRS	0.628±0.067	0.727±0.065	0.627±0.069	0.023±0.005	0.045±0.010	0.012±0.002	0.731±0.020
	Per Diagnosis	0.653±0.058	0.717±0.051	0.653±0.060	0.025±0.004	0.047±0.008	0.012±0.002	0.742±0.021
	QRISK3	0.618±0.069	0.737±0.083	0.617±0.071	0.023±0.004	0.044±0.008	0.012±0.002	0.718±0.019
	SCORE2	0.607±0.068	0.720±0.062	0.606±0.069	0.022±0.003	0.042±0.006	0.012±0.002	0.712±0.017
	AHA/ASCVD	0.625±0.074	0.693±0.066	0.624±0.076	0.022±0.005	0.043±0.009	0.012±0.002	0.704±0.020
	FGCRS	0.657±0.160	0.593±0.161	0.657±0.164	0.022±0.006	0.042±0.011	0.012±0.002	0.680±0.017
Hemorrhagic stroke	UKCRP	0.557±0.153	0. 705±0.135	0.557±0.154	0.005±0.001	0.011±0.002	0.004±0.001	0.644±0.026
	UKCRP+PRS	0.571±0.178	0.626±0.170	0.571±0.179	0.006±0.0001	0.011±0.003	0.004±0.001	0.646±0.026
	Per Diagnosis	0.625±0.108	0.628±0.098	0.625±0.0109	0.006±0.002	0.012 ±0.003	0.004±0.001	0.659±0.031
	QRISK3	0.531±0.091	0.705±0.098	0.530±0.092	0.005±0.001	0.011±0.002	0.004±0.001	0.642±0.019
	SCORE2	0.551±0.092	0.679±0.095	0.550±0.093	0.005±0.001	0.011±0.002	0.004±0.001	0.638±0.028
	AHA/ASCVD	0.535±0.112	0.691±0.113	0.534±0.113	0.005±0.001	0.011±0.002	0.004±0.001	0.636±0.028
	FGCRS	0.368±0.222	0.755±0.231	0.368±0.224	0.004±0.001	0.009±0.002	0.004±0.001	0.589±0.025

Binarization cut-off was determined based on the achievement of the largest Youden index (Youden index = sensitivity + specificity - 1).

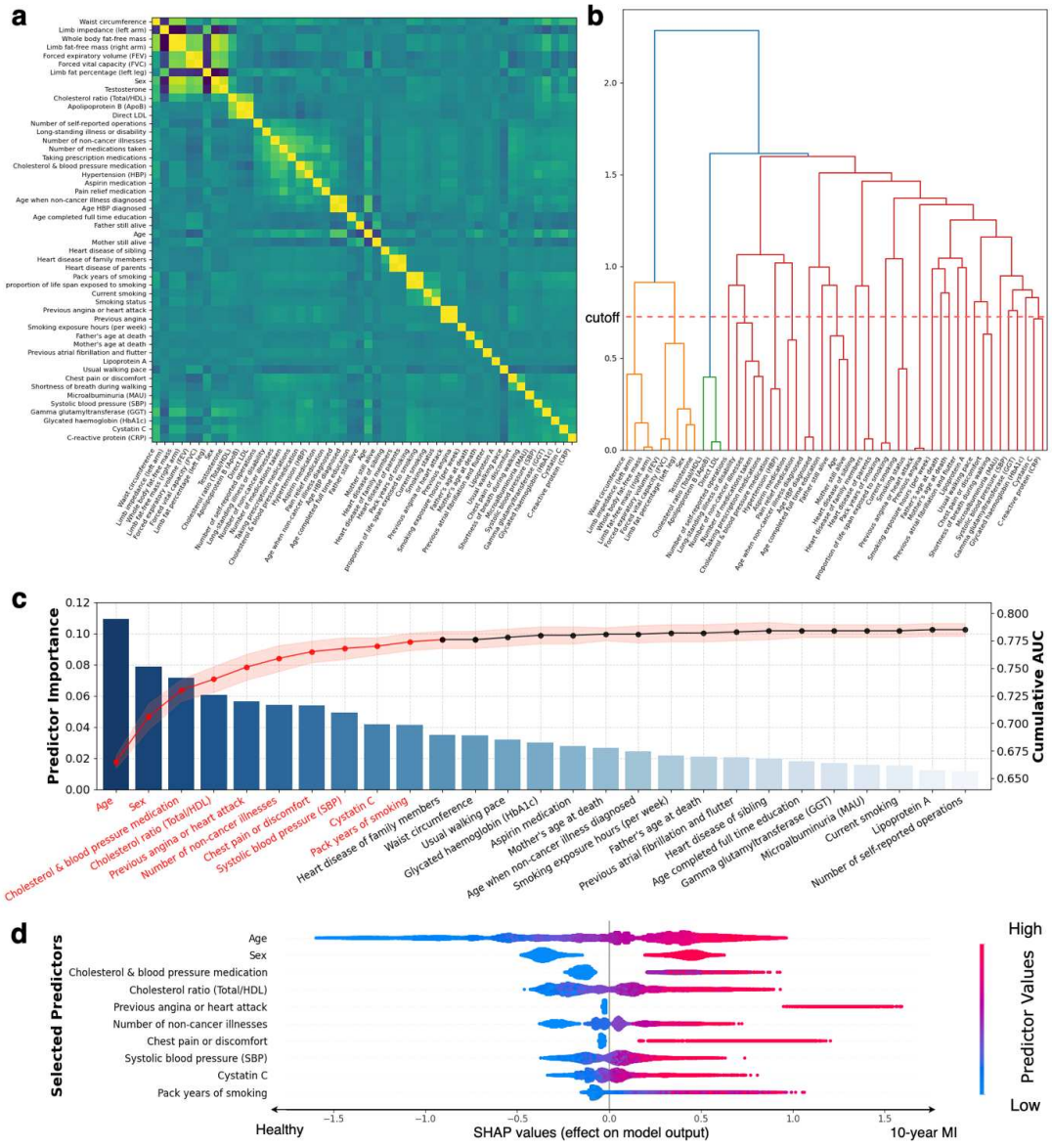
Abbreviations: PRS = polygenic risk score, SCORE2 = Systematic Coronary Risk Evaluation 2, AHA/ASCVD = American Heart Association/Atherosclerotic Cardiovascular Disease, FGCRS = Framingham Cardiovascular Risk Score

eFigure1: Correlations heatmap and clustering dendrogram of the top-50 candidate predictors on 10-year incident cardiovascular disease



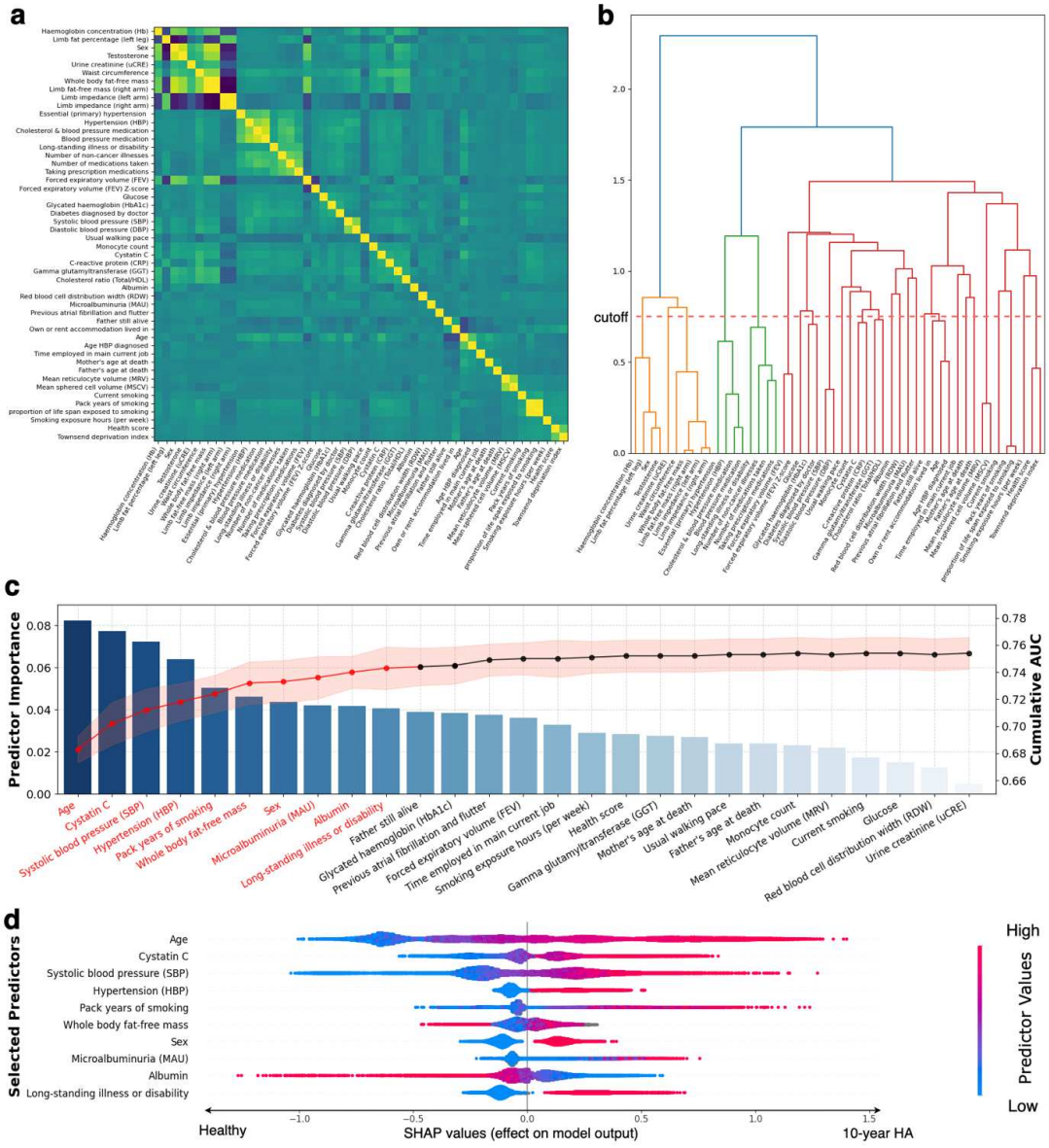
(a) Heatmap of Spearman rank-order correlations between each pair of the top-50 candidate predictors of modeling on 10-year incident cardiovascular disease; (b) dendrogram of hierarchical clustering based on calculated correlations. The horizontal dash line, 0.75, was the cutoff of clusters, and only one predictor was chosen within each cluster (grouped predictors under a threshold of 0.75).

eFigure2: Predictor selection and interpretation on 10-year incident myocardial infarction



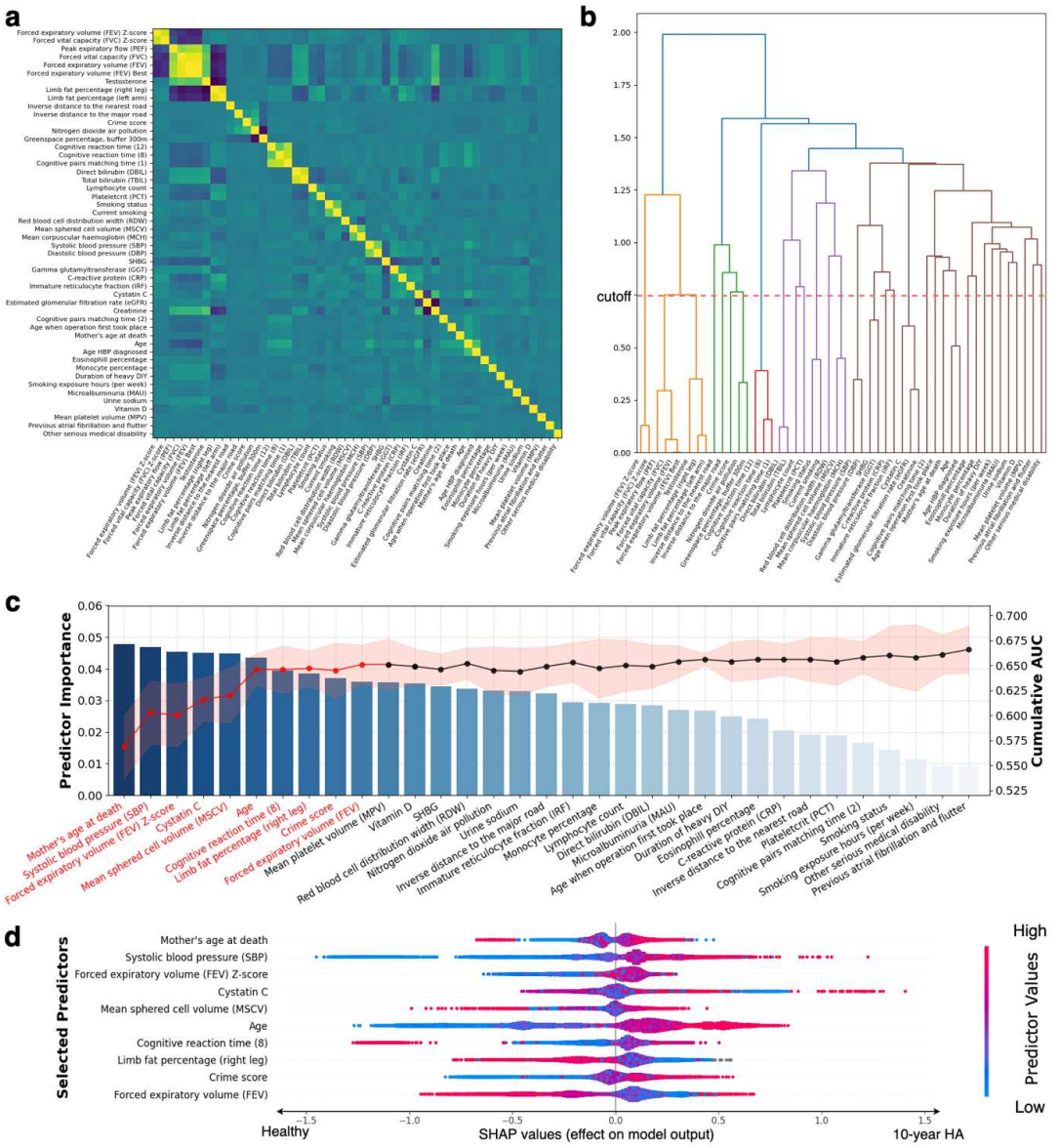
(a) Heatmap of Spearman rank-order correlations between each pair of the top-50 candidate predictors of modeling on 10-year incident myocardial infarction; (b) dendrogram of hierarchical clustering based on calculated correlations. The horizontal dash line, 0.75, was the cutoff of clusters, and only one predictor was chosen within each cluster (grouped predictors under a threshold of 0.75); (c) sequential forward selection of pre-selected candidate predictors; (d) SHAP visualization plot of the selected predictors.

eFigure3: Predictor selection and interpretation on 10-year incident ischemic stroke



(a) Heatmap of Spearman rank-order correlations between each pair of the top-50 candidate predictors of modeling on 10-year incident ischemic stroke; (b) dendrogram of hierarchical clustering based on calculated correlations. The horizontal dash line, 0.75, was the cutoff of clusters, and only one predictor was chosen within each cluster (grouped predictors under a threshold of 0.75); (c) sequential forward selection of pre-selected candidate predictors; (d) SHAP visualization plot of the selected predictors.

eFigure4: Predictor selection and interpretation on 10-year incident hemorrhagic stroke



(a) Heatmap of Spearman rank-order correlations between each pair of the top-50 candidate predictors of modeling on 10-year incident hemorrhagic stroke; (b) dendrogram of hierarchical clustering based on calculated correlations. The horizontal dash line, 0.75, was the cutoff of clusters, and only one predictor was chosen within each cluster (grouped predictors under a threshold of 0.75); (c) sequential forward selection of pre-selected candidate predictors; (d) SHAP visualization plot of the selected predictors.