OXFORD

Briefings in Bioinformatics, 2024, **25(6)**, bbae479

https://doi.org/10.1093/bib/bbae479

Case Study

AESurv: autoencoder survival analysis for accurate early prediction of coronary heart disease

Yike Shen[1,2], Arce Domingo-Relloso[3,4,5], Allison Kupsco[2], Marianthi-Anna Kioumourtzoglou[2], Maria Tellez-Plaza[3], Jason G. Umans[6], Amanda M. Fretts[7], Ying Zhang[8], Peter F. Schnatz[9], Ramon Casanova[10], Lisa Warsinger Martin[11], Steve Horvath[12,13], JoAnn E. Manson[14], Shelley A. Cole[15], Haotian Wu[2], Eric A. Whitsel[16], Andrea A. Baccarelli[2,17], Ana Navas-Acien[2], Feng Gao[18,19,*]

[1]Department of Earth and Environmental Sciences, University of Texas at Arlington, 500 Yates Street, Arlington, TX, 76019, USA
[2]Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, 722 West 168th Street, New York, NY, 10032, USA
[3]Department of Chronic Diseases Epidemiology, National Center for Epidemiology, Carlos III Health Institute, C. de Melchor Fernández Almagro, 5, Fuencarral-El Pardo, 5, Madrid, 28029, Spain
[4]Department of Statistics and Operations Research, University of Valencia, Carrer del Dr. Moliner, 50, Valencia, 46100, Spain
[5]Department of Biostatistics, Columbia University Mailman School of Public Health, 722 West 168th Street, New York, NY, 10032, USA
[6]Department of Medicine, Georgetown-Howard Universities Center for Clinical and Translational Science, 4000 Reservoir Road NW, Washington, DC, 20007, USA
[7]Department of Epidemiology, University of Washington, 3980 15th Ave NE, Seattle, WA, 98195, USA
[8]Center for American Indian Health Research, Department of Biostatistics and Epidemiology, The University of Oklahoma Health Sciences Center, 801 N.E. 13th Street, Oklahoma City, OK, 73104, USA
[9]Department of OB/GYN and Internal Medicine, Reading Hospital/Tower Health & Drexel University, 301 S 7th Ave, West Reading, PA, 19611, USA
[10]Department of Biostatistics and Data Science, Wake Forest University School of Medicine, 475 Vine St, Winston Salem, NC, 27101, USA
[11]Department of Medicine, Division of Cardiology, George Washington University, 2300 Eye Street, NW, Washington, DC, 20037, USA
[12]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles (UCLA), 695 Charles E. Young Drive South, Los Angeles, CA, 90095, USA
[13]Altos Lab Inc, Granta Park, Little Abington, Cambridge, CB21 6GQ, United Kingdom
[14]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 900 Commonwealth Ave, Boston, MA, 02215, USA
[15]Population Health Program, Texas Biomedical Research Institute, 8715 W. Military Dr., San Antonio, TX, 78227, USA
[16]Department of Epidemiology, Gillings School of Global Public Health and Department of Medicine, School of Medicine, University of North Carolina at Chapel Hill, 135 Dauer Drive, Chapel Hill, NC, 27599, USA
[17]Harvard T.H. Chan School of Public Health, Harvard University, 677 Huntington Avenue, Boston, MA, 02115, USA
[18]Department of Environmental Health Sciences, Fielding School of Public Health, University of California Los Angeles (UCLA), 650 Charles E. Young Drive South, Los Angeles, CA, 90095, USA
[19]Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles (UCLA), 650 Charles E. Young Drive South, Los Angeles, CA, 90095, USA

[*]Corresponding author. Feng Gao, Department of Environmental Health Sciences, Fielding School of Public Health and Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, 650 Charles E. Young Drive South, Room 56-070C (CHS), University of California Los Angeles (UCLA), Los Angeles, CA, 90095, USA. Tel.: 5175755686; E-mail: gaofeng@ucla.edu

Abstract

Coronary heart disease (CHD) is one of the leading causes of mortality and morbidity in the United States. Accurate time-to-event CHD prediction models with high-dimensional DNA methylation and clinical features may assist with early prediction and intervention strategies. We developed a state-of-the-art deep learning autoencoder survival analysis model (AESurv) to effectively analyze high-dimensional blood DNA methylation features and traditional clinical risk factors by learning low-dimensional representation of participants for time-to-event CHD prediction. We demonstrated the utility of our model in two cohort studies: *the Strong Heart Study cohort (SHS)*, a prospective cohort studying cardiovascular disease and its risk factors among American Indians adults; the *Women's Health Initiative (WHI)*, a prospective cohort study including randomized clinical trials and observational study to improve postmenopausal women's health with one of the main focuses on cardiovascular disease. Our AESurv model effectively learned participant representations in low-dimensional latent space and achieved better model performance (concordance index-C index of 0.864 ± 0.009 and time-to-event mean area under the receiver operating characteristic curve-AUROC of 0.905 ± 0.009) than other survival analysis models (Cox proportional hazard, Cox proportional hazard deep neural network survival analysis, random survival forest, and gradient boosting survival analysis models) in the SHS. We further validated the AESurv model in WHI and also achieved the best model performance. The AESurv model can be used for accurate CHD prediction and assist health care professionals and patients to perform early intervention strategies. We suggest using AESurv model for future time-to-event CHD prediction based on DNA methylation features.

Keywords: autoencoder survival analysis; deep learning; coronary heart disease; cohort studies; epigenetics

Received: March 22, 2024. **Revised:** August 17, 2024. **Accepted:** September 12, 2024

Introduction

Coronary heart disease (CHD) is one of the leading causes of mortality and morbidity in the United States [1–4]. Accurate time-to-event CHD prediction models are needed to provide early prediction and assist decisions regarding implementation of intervention strategies. Clinical features and traditional risk factors such as smoking and blood pressure have been widely used for CHD prediction [1, 5]. On the other hand, various studies have indicated that differential DNA methylation, a reversible epigenetic mechanism that adds methyl groups to cytosine and thus modifies genome function, is associated with CHD [6–12]. Furthermore, previous epidemiological studies have identified that variations in DNA methylation are associated with CHD risk factors such as air pollution [13], cardiotoxic metal exposures [14, 15], smoking [1, 16], hypertension [17, 18], obesity [19], and type-2 diabetes [10, 12, 13, 16–19]. For example, Turunen et al., [20] found that reductions in DNA methylation have been linked to atherosclerosis in various tissues, which is a chronic disease that may contribute to cardiovascular disease morbidity and mortality [12, 20]. Indeed, DNA methylation can reflect the effects of cumulative cardiovascular risk factor exposures on epigenetics and provide rich information for CHD prediction as biomarkers.

While DNA methylation data have been previously used towards CHD prediction, existing models usually focused on a few selected DNA methylation sites (CpG sites) together with other features such as single nucleotide polymorphisms (SNPs) or clinical risk factors to predict binary CHD [7–9]. However, binary CHD prediction does not account for the time-to-event nature of CHD development. This can be better addressed by survival analysis, which studies the time of certain CHD event (e.g., CHD death, first occurrence of clinical myocardial infarction, etc). Classic Cox proportional hazard model (CPH) estimates log hazard through linear model, thus may not capture nonlinearity within the data. Therefore, DeepSurv, a fully connected neural network based Cox survival analysis model, has been proposed to account for non-linear relationships [21]. In addition, the current cutting-edge Illumina Methylation Array measures more than 850 000 or 450 000 epigenome-wide DNA methylation sites [22, 23], which facilitates the discovery of hundreds of significant DNA methylation sites related to CHD through epigenome wide association studies. These newly identified significant sites form high-dimensional data [6, 24, 25]. For example, in a previous study, 635 noteworthy CpG sites were discovered in the Strong Heart Study (SHS), while 398 noteworthy CpG sites were identified in the Women's Health Initiative (WHI) [6]. In the same study, the Framingham Heart Study and Atherosclerosis Risk in Communities Study discovered 698 and 2092 noteworthy CpG sites, respectively [6]. The study found 29 common CpG sites across at least four cohorts [6]. While these identified noteworthy sites help reduce the dimension of DNA methylation data for downstream analysis, the hundreds of CpG sites are still high-dimensional. Therefore, it is desirable to learn representation of the participants from the high-dimensional DNA methylation data to improve prediction of time-to-event CHD. One way of learning representations of high-dimensional data is through an autoencoder model. Autoencoder is an unsupervised deep learning model that learns low dimensional representations (embeddings) from high dimensional data [26]. When combined with a CPH model, the autoencoder can learn embeddings of DNA methylation and clinical data and leverage it towards time-to-event survival analysis.

In this study, we developed a novel deep learning autoencoder survival analysis (AESurv) model. Our AESurv model tackles non-linear relationships and effectively learns low dimensional latent space representation of participants from high dimensional input DNA methylation data and clinical features. The developed model utilized noteworthy DNA methylation CpG sites generated for the American Indian communities in the Strong Heart Study in combination with clinical features. We validated our model in the Women's Health Initiative with noteworthy CpG sites alone. Our state-of-the-art prediction of time-to-event CHD can serve as early signals of CHD risk in baseline healthy individuals and assist in the development of early intervention strategies.

Methods

Schematic workflow of our study is shown in Fig. 1. Clinical and DNA methylation data were first obtained from population cohort studies. We then utilized the proposed AESurv model to learn low-dimensional representations from the combined clinical and DNA methylation features. We further compared the model performances towards predicting time-to-event CHD with other survival machine learning models including cox proportional hazard model, cox proportional hazards deep neural network model, random survival forest (RSF), and gradient boosted survival analysis.

Study population

Strong Heart Study. The SHS cohort was established to study the disproportionally high burden of CHD in American Indian communities in the Southwest and the Great Plains. Incident CHD in the SHS cohort in this study includes both fatal (sudden death due to CHD and first occurrence of definite fatal myocardial infarction) and non-fatal (definite non-fatal CHD and non-fatal myocardial infarction) events. The current study was conducted on 2321 participants with 16.45 years of follow-up and 749 incident CHD events during follow-up (338 are fatal and 411 are non-fatal). Our participants met the same inclusion criteria as described in Navas-Acien [6], with no coronary heart disease or missing data for risk factors of cardiovascular disease at baseline, but with available blood DNA methylation measures. The SHS is a participatory based study working in partnership with tribal communities in the Southwest, the Northern plains, and the Southern Plains. Participating tribal communities and institutional review boards (IRBs) of participating institutions and the respective area Indian Health Service approved the protocol. Informed consent was provided by all participants.

Women's Health Initiative. WHI is a prospective cohort study including randomized clinical trials (CTs) and observational study (OS) to improve postmenopausal women's health with one of the main focuses on cardiovascular disease. WHI CHD is defined as the definite silent myocardial infarction, first occurrence of clinical myocardial infarction, or a death due to possible or definite CHD in main WHI, extension 1, and extension 2 CT and OS population. The WHI participants in this study were drawn from Broad Agency Announcement 23 (WHI-BAA23, an incident CHD case–control study) with available blood DNA methylation data [27]. Few missing CpG sites that did not pass quality control were imputed with k-nearest neighbors algorithm. WHI participants who were included in this study were free of CHD at baseline and had blood DNA methylation measures, resulting in a total of 2107 participants with average of 17.33 years of follow-up and 706 CHD events during follow-up. WHI participants in our

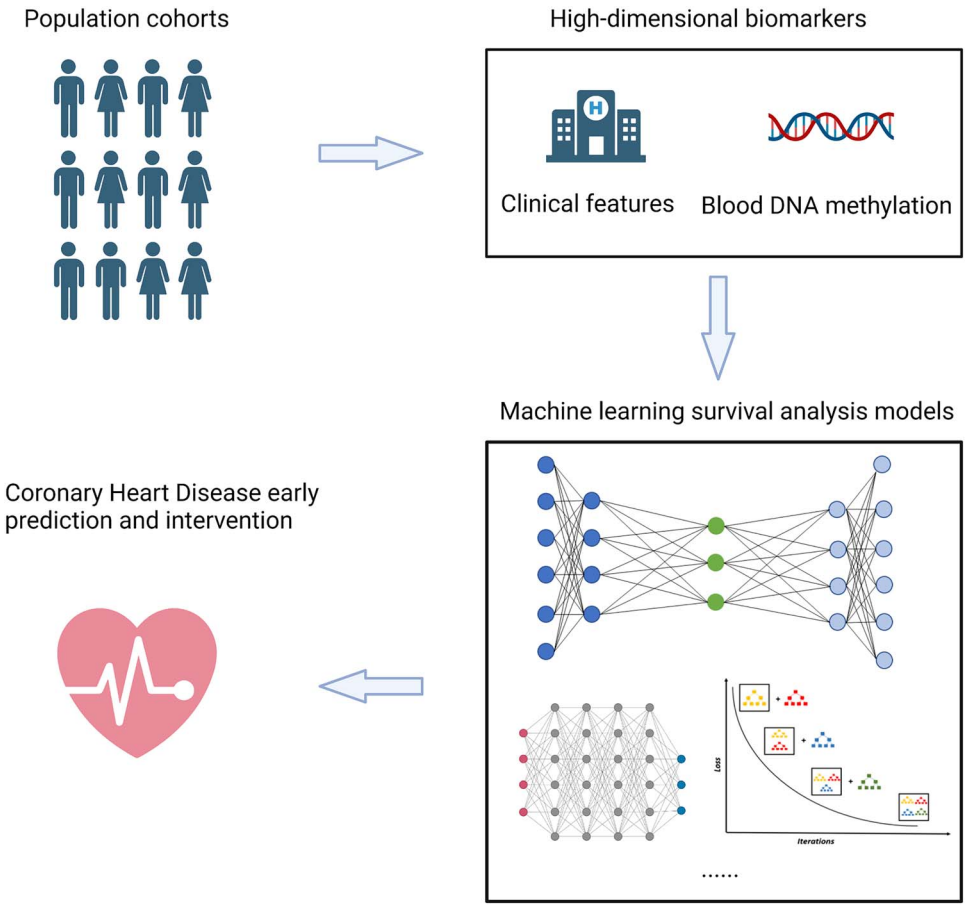

Figure 1. Schematic workflow using machine learning survival analysis models with high-dimensional biomarkers from population cohorts.

study had an average age of 64.47 at screening. The majority of race/ethnicity in WHI population are White (58%) and Black (27%). All WHI study protocols were approved by IRBs at multiple academic institutions and participants. Informed consent was provided by all participants.

DNA methylation data

The MethylationEPIC BeadChip (Illumina 850 K) was used to measure Blood DNA methylation for SHS, while the Human-Methylation450 BeadChip (covering 450 K CpG sites) was used for measurements in WHI. We followed the same quality check and inclusion criteria for DNA methylation sites as described in Navas-Acien [6]. The M-values (SHS) and beta value (WHI) of CpG sites were screened through an elastic-net penalized Cox proportional hazard model to identify noteworthy CpG sites associated with time-to-event CHD. The resulting 635 noteworthy CpG sites in SHS and 398 noteworthy CpG sites in WHI from our previous epigenome wide association studies [6] were used as our input DNA methylation features in this study.

Clinical features for Strong Heart Study

Continuous variables that are used as input features were age (years), BMI (kg/m^2), systolic blood pressure (mmHg), high-density lipoprotein cholesterol (mg/dL), low-density lipoprotein cholesterol (mg/dL). Categorical features included sex (male or female), type 2 diabetes (yes or no), hypertension treatment (yes or no), smoking status (current, former, or never), albuminuria status (normal, microalbuminuria, or macroalbuminuria), study center (Oklahoma, Arizona, or South Dakota and North Dakota), and proportions of CD4T, CD8T, NK, B cells, and monocytes. For simplicity, we call these variables 'clinical features' throughout. One-hot encoding was utilized for categorical features (i.e., created dummy variables) to generate the final set of 25 clinical features (Table S1). Clinical features were only used in SHS analyses.

Autoencoder survival analysis model

We developed a deep learning AESurv model utilizing a supervised autoencoder combined with the average negative log partial likelihood loss function from CPH. Originally, autoencoder is an unsupervised deep learning method that consists of an encoder and a decoder (Fig. 2) [26, 28]. The encoder takes high-dimension input features x_i and reduces x_i to lower-dimension embeddings (representative features), whereas the decoder outputs the reconstructed feature $\hat{x}_i$. Here we adapt the unsupervised autoencoder model into a supervised one by using the learned embedding to predict the log hazard ratio $\hat{h}_{w(x_i)}$ in log hazard function. Specifically, the network output $\hat{h}_{w(x_i)}$ estimates the log-risk function in the Cox model which enables us to conduct time-to-event prediction [21]. To train our AESurv model, we combined the loss function for autoencoder to reconstruct the original features (reconstruction loss that measured the differences between the original features and reconstructed features) and the average negative log partial likelihood loss function from CPH (Cox loss). We also added L2 regularization ($\lambda\ ||w||_2^2$) to avoid overfitting. We define $N_{E=1}$ as the number of participants with CHD and the set

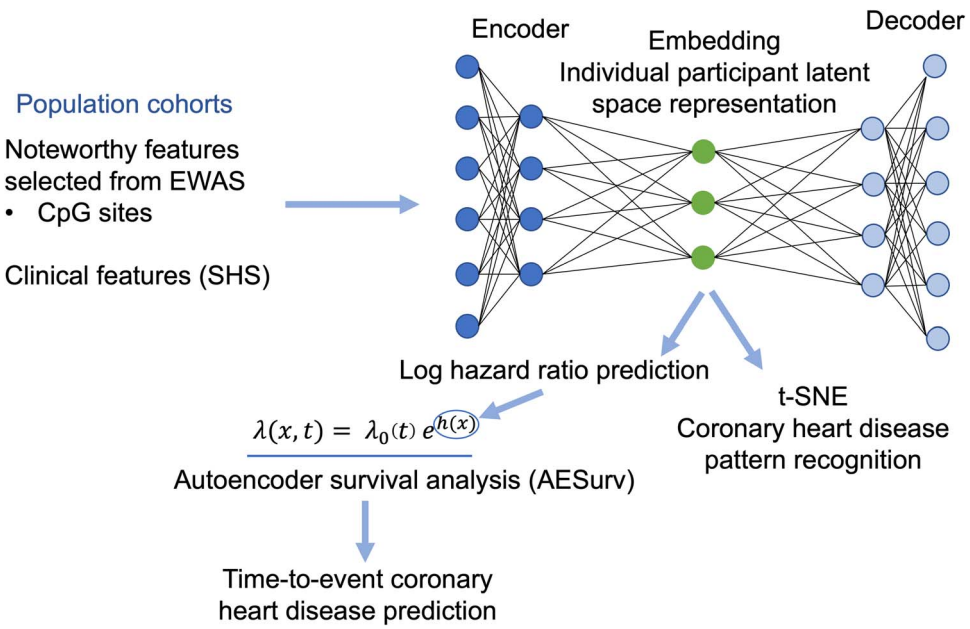

Figure 2. Conceptual of autoencoder survival analysis model (AESurv). Epigenome wide association study (EWAS); strong heart study (SHS).

of participants still at risk of failure at time t as $R(t) = \{i:T_i \geq t\}$. The full loss function is defined as:

$$loss_{AESurv} = \underbrace{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}_{Reconstruction\ loss} - \underbrace{\frac{1}{N_{E=1}} \sum \left(\hat{h}_w(x_i) - \log \sum_{j \in R(T_i)} e^{\hat{h}_w(x_j)} \right)}_{Cox\ loss} + \lambda\, ||w||_2^{\,2} \quad (1)$$

The autoencoder encoding-decoding process effectively learns low-dimensional participant representations. Meanwhile, the prediction process enables the information related to CHD to be kept in the embeddings. Our model enables an autoencoder survival analysis for both feature representation and time-to-event prediction. The 635 DNA methylation and 25 clinical features from SHS were input together into our autoencoder model to learn participant low dimensional representations/embeddings.

Specifically, our AESurv comprises three parts: an encoder that learns a latent space representation of each participant based on the initial 660 input features (635 DNA methylation features and 25 clinical features), a decoder that rebuilds the input features (Fig. 2), and a linear prediction component of the log hazard ratio. The number of nodes for our model in three hidden layers are 256, 32, and 256, respectively. The layer with 32 neurons is the embedding layer. Various embedding sizes, including 2, 16, 32, 64, 128, and 256, were compared to select the optimal model structures. ReLU activation function was used [29–31]. The hyperparameter set was chosen according to the best model performance from the validation set. The AESurv model was trained with L2 regularization weight of 0.0001, the Adam optimizer [32], batch size of 128, dropout rate of 0.5, and learning rate of 0.0001. Early stopping (epoch stopped when performance of the model on validation dataset started to decrease) was used to prevent overfitting. We repeated the same process in WHI with 398 input DNA methylation features.

Model interpretation

The learned latent space embeddings were first visualized using t-SNE, which has been widely used for visualizing high-dimensional data [33]. T-SNE can embed the local structure of the data into low dimensional spaces and reveal patterns in the data. T-SNE minimizes the Kullback–Leibler divergence between the original high-dimensional data and the low-dimensional embedding. Here we utilized t-SNE to discover CHD patterns in the data. In the t-SNE plot, participants were clustered based on the learned embeddings of DNA methylation and clinical features. The results were colored with or without CHD to visualize the effectiveness of learned embeddings.

Survival analysis with other models

The Cox proportional hazards model is a regression model that is used to investigate the simultaneous effect of risk factors on survival time [34]. It assumes proportional hazards and linear relationships. Splines can be used together with CPH model to incorporate nonlinear relationships [35]. Katzman [21] developed a DeepSurv model that used a non-linear neural network based log hazard ratio in the CPH log hazard function. Additionally, RSF is an ensemble method for right-censored survival data, which do not assume proportional hazards and takes non-linear effects into consideration [36]. Gradient boosted survival analysis (GBRTS) is an additive model that minimized the partial likelihood loss by adding regression trees [37]. It combines the multiple base learners' predictions to achieve a better overall model. Hyperparameters for CPH, DeepSurv, RSF, and GBRTS were tuned with five-fold cross validation. We used same CHD endpoint for all models for comparability.

We used five-fold cross validation in our supervised machine learning model. To summarize, we first randomly shuffle and split the dataset into five equal groups, where each of the group was used as a test set and the other four groups were used as training data. To identify the best parameter combination, the training data were further split into 10% for validation and 70% for training. Therefore, the dataset was split into individually held 20% test, 70% training, and 10% validation. The best parameter sets were selected based on the model performance from the validation set. We then used the best parameter combination to build the model and test on the unseen test set. We averaged the

Table 1. Embedding size comparison of AESurv models in the strong heart study.

Embedding size	C-index	mean AUROC
2	0.862 ± 0.014	0.903 ± 0.012
16	0.862 ± 0.012	0.903 ± 0.014
32	0.864 ± 0.009	0.905 ± 0.009
64	0.861 ± 0.016	0.902 ± 0.015
128	0.860 ± 0.01	0.901 ± 0.009
256	0.859 ± 0.013	0.900 ± 0.014
512	0.857 ± 0.011	0.898 ± 0.011

five-split performance results to get the final prediction results. Therefore, each sample is given the opportunity to be tested once and used to train the model four times. The five-fold cross validation procedure were then conducted five times to obtain their average results. We reported the average prediction accuracy and standard deviation (± SD). We computed Concordance index (C-index), time-to-event mean Area Under the Receiver Operating Characteristic curve (AUROC), and time-dependent AUROC as measures of predictive accuracy. C-index is a generalization of the AUROC curve to consider censored data. It reflects the model's ability to accurately rank survival times [38, 39]. It calculates as the number of concordant pairs divided by the sum of number of concordant pairs and discordant pairs. The higher the AUROC and C-index, the better model performance. Time-dependent AUROC determines how well a model can perform at certain time points given the disease status at that time point. Time-to-event mean AUROC is calculated as the average of all time-dependent AUROC. Our AESurv model was implemented using Pytorch (Version 1.10.1) and [37]. We compared AESurv results with other machine learning and deep learning models, including CPH, DeepSurv, RSF, and GBRTS. We implemented the other models with Scikit-survival (Version 0.16.0), PyTorch (Version 1.10.1)), and R survival package (Version 3.5–7). The de-identified code is available at https://github.com/YikeShen/AESurv.

Results and discussion

Exploration of CHD patterns with participant representations in the latent space

Our AESurv model effectively learned the participant representation from the DNA methylation and clinical features in a low-dimensional latent space that can be used towards CHD prediction. We first compared the model performances of our AESurv model with different embedding sizes: 2, 16, 32, 64, 128, 256, and 512 (Table 1). Our results showed that the AESurv model is robust to the choice of embedding sizes with embedding size 32 achieved the highest model performance (C-index = 0.864 ± 0.009, mean AUROC = 0.905 ± 0.009). We thus selected 32 as the optimal embedding size.

We then visualized the raw input features of the training dataset (Fig. 3a), the learned embeddings of training (Fig. 3b), and test dataset (Fig. 3c) under their t-SNE coordinates colored by with or without CHD. We randomly selected one of the five splits in one repeat for visualization. The purple color refers to participants without CHD and yellow color refers to participants with CHD. Our results showed that the CHD patterns can be reflected by the participant embeddings. There were clear clusters of participants with or without CHD in the test dataset using learned participant representations (Fig. 3). The ability to reflect CHD diagnoses based on learned DNA methylation and clinical feature embeddings suggests that our AESurv method effectively learns low-dimensional latent space representations of high-dimensional DNA methylation and clinical features.

Table 2. Time-to-event model performance of the strong heart study. DNA methylation features were selected through elastic-net penalized cox regression from epigenome-wide DNA methylation array (850 K). Clinical features are listed in the materials and methods. Numbers are prediction score ± standard deviation of five repeat runs. AESurv = deep autoencoder survival analysis model, DeepSur = Cox proportional hazards deep neural network model, RSF = random survival forest model, GBRTS = gradient boosted survival analysis model, CPH = Cox proportional hazard model.

SHS DNA methylation features (635) + Clinical features (25)		
Model	**C-index**	**mean AUROC**
AESurv	0.864 ± 0.009	0.905 ± 0.009
DeepSurv	0.855 ± 0.013	0.897 ± 0.012
RSF	0.683 ± 0.023	0.719 ± 0.024
GBRTS	0.710 ± 0.027	0.757 ± 0.029
CPH	0.855 ± 0.014	0.898 ± 0.015
DNA methylation features (635)		
AESurv	0.853 ± 0.01	0.893 ± 0.009
DeepSurv	0.845 ± 0.013	0.885 ± 0.011
RSF	0.645 ± 0.027	0.670 ± 0.031
GBRTS	0.689 ± 0.019	0.730 ± 0.021
CPH	0.845 ± 0.013	0.887 ± 0.013
Clinical features (25)		
AESurv	0.706 ± 0.017	0.759 ± 0.02
DeepSurv	0.708 ± 0.015	0.761 ± 0.016
RSF	0.694 ± 0.019	0.742 ± 0.024
GBRTS	0.679 ± 0.014	0.729 ± 0.015
CPH	0.713 ± 0.014	0.767 ± 0.017

Prediction of time-to-event CHD

We utilized our AESurv model to predict the time-to-event CHD combining the selected DNA methylation and clinical features. Additionally, we compared the performance of other commonly used survival analysis models—DeepSurv, CPH, RSF, GBRTS. Our AESurv model achieved the best model performance compared to all other survival analysis models with the highest C-index of 0.864 ± 0.009 and time-to-event mean AUROC of 0.905 ± 0.009 (Table 2). Time-dependent AUROC of AESurv model had better performance than tree-based RSF and GBRTS model (Fig. 4). The C-index gives a rank-based prediction score while the time-dependent AUROC shows survival function with the mean AUROC averaged across time. CPH and DeepSurv models had similar performance with the similar mean AUROC (CPH = 0.898 ± 0.015, DeepSurv = 0.897 ± 0.012) and C-index (CPH = 0.855 ± 0.014, DeepSurv = 0.855 ± 0.013) (Table 2). DNA methylation of adjacent CpGs within the same CpG island or regulatory region may be co-regulated, leading to the detection of multiple co-regulated CpGs in a single EWAS. Some distant CpGs could be interrelated as well [40, 41]. Inclusion of all these CpGs within models increases the dimensionality of the data, leading to potentially redundant information. Our AESurv model successfully reduced the dimensionality and learned the latent space representation of DNA methylation features. The representative features in the low dimensional latent space had more predictive power than the raw input features.

The performances of the tree-based survival analysis models were worse than AESurv, DeepSurv, and CPH, with the C-index being 0.683 ± 0.023 and mean AUROC of 0.719 ± 0.024 for RSF and C-index of 0.710 ± 0.027 and mean AUROC of 0.757 ± 0.029 for GBRTS (Table 2). Previous studies used tree-based models

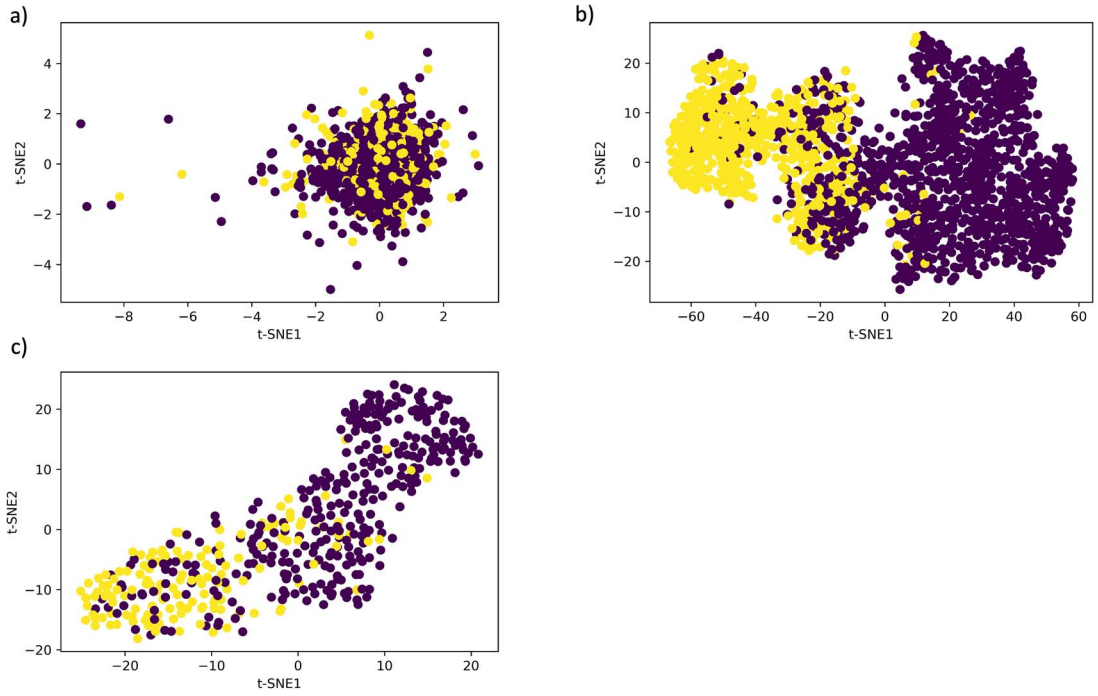

Figure 3. Visualization of DNA methylation and clinical feature embeddings in t-SNE coordinates in the strong heart study. (a) t-SNE of raw EWAS selected features; (b) t-SNE of autoencoder learned embeddings in training dataset; (c) t-SNE of autoencoder learned embeddings in test dataset.

with selected DNA methylation and/or SNPs to predict binary CHD outcomes. Dogan [7] achieved an AUROC of 0.78 in the Framingham Heart Study with random forest classification on binary CHD outcomes. Cugliari [9] predicted the fatal and non-fatal cardiovascular disease events using random forest model in an Italian cardiovascular cohort (EPICOR cohort) with AUROC of 0.74. However, predicting time-to-event CHD can better account for time component in CHD development and could be more challenging.

Ablation study of DNA methylation and clinical features

To further understand the contribution of DNA methylation and clinical features to predict time-to-event CHD, we tested model performance using only clinical features and only DNA methylation features. The results of CHD prediction using 635 DNA methylation features alone had slightly worse performance (AUROC = 0.853 ± 0.01) than using both DNA methylation and clinical features (AUROC = 0.864 ± 0.009) (Table 2). Using 25 clinical features alone has the lowest prediction accuracy than DNA methylation alone and the combination of DNA methylation and clinical features (Table 2). Previous studies also found clinical features and traditional risk factors may have substantial residual risks in CHD prediction [1, 5, 7, 8, 24]. For example, Wilson [1] incorporated clinical features including blood pressure and cholesterol in CHD prediction and noted that other factors, may also contribute to CHD prediction but are not included in the model due to data availability. Since clinical features and traditional risk factors are usually selected based on prior knowledge, it would be almost impossible to capture all potential CHD risk factors. On the other hand, the effects of cumulative cardiovascular risk factor exposures can be reflected by DNA methylation. By utilizing rich DNA methylation data, researchers and clinicians now have more power in accurate time-to-event CHD prediction to assist in early intervention of patients with high risk for CHD.

Table 3. Time-to-event model performance of Women's health initiative. DNA methylation features were selected through elastic-net penalized cox regression from epigenome-wide DNA methylation array (450 K). Clinical features are listed in the materials and methods. Numbers are prediction score ± standard deviation of five repeat runs. AESurv = deep autoencoder survival analysis model, DeepSur = Cox proportional hazards deep neural network model, RSF = random survival forest model, GBRTS = gradient boosted survival analysis model, CPH = Cox proportional hazard model.

WHI DNA methylation features (398)	
Model	**C-index**
AESurv	0.752 ± 0.019
DeepSurv	0.741 ± 0.023
RSF	0.457 ± 0.014
GBRTS	0.457 ± 0.014
CPH	0.725 ± 0.021
CPH-Splines	0.624 ± 0.016

Validation in Women's Health Initiative

To further investigate our model applicability across different cohorts, we tested our AESurv in the Women's Health Initiative. From the SHS ablation study, DNA methylation features alone have good performance and adding clinical features only incrementally increased prediction accuracy. Therefore, we focused on WHI with DNA methylation features. Similarly, AESurv had the best performance (C-index = 0.752 ± 0.019) than DeepSurv (C-index = 0.741 ± 0.023), CPH (C-index = 0.725 ± 0.021), and CPH with splines (C-index = 0.624 ± 0.016) (Table 3). Similarly, tree-based survival analysis models had the worst performance of 0.457 ± 0.014 for both RSF and GBRTS (Table 3).

We note that there are several differences between SHS and WHI cohorts that may contribute to different prediction accuracy in two cohorts. For example, the noteworthy CpG sites from SHS were obtained from 850 K DNA methylation array, while the CpG

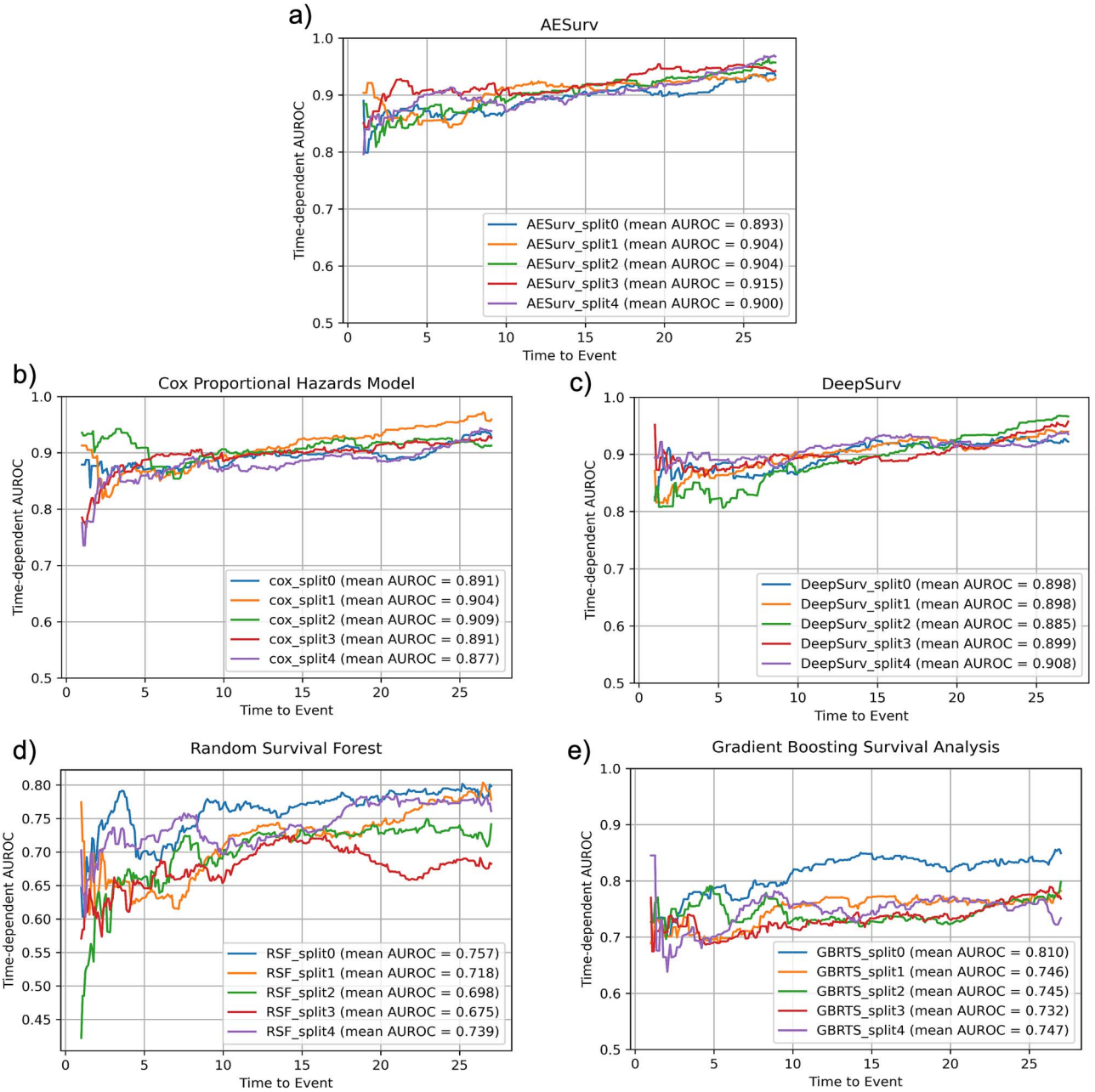

Figure 4. Strong heart study time-dependent AUROC (one randomly selected repeat (total 5). AESurv = deep autoencoder survival analysis model, DeepSurv = Cox proportional hazards deep neural network model, RSF = random survival forest model, GBRTS = gradient boosted survival analysis model, CPH = Cox proportional hazard model.

sites from WHI were obtained from 450 K DNA methylation array. Additionally, WHI is a female cohort, whereas SHS includes both male and female participants. However, our ablation results from SHS showed that adding gender as a feature only incrementally increased prediction accuracy. Finally, our AESurv model consistently performs better than other survival analysis models in different population cohorts, highlighting the applicability of our model for coronary heart disease prediction.

Limitations and future directions

In this study, we used two distinct populations, an American Indian population (SHS) and a female population (WHI), with over 2000 participants in each cohort. While our AESurv model consistently achieved better results in both populations, the performance of our model in other populations is yet to be explored. In the future, we could expand to larger populations or consortia to increase generalizability. We could also predict different subcategories of CHD, such as fatal and non-fatal CHD, to provide more nuanced predictions. Finally, we could further adapt our AESurv model to directly learn from 850 K or 450 K DNA methylation array data.

Conclusion

We developed a novel AESurv model to analyze high-dimensional DNA methylation features and predict time-to-event CHD, which can contribute to early prediction and clinical intervention of CHD. Our model achieved the state-of-the-art prediction accuracy of CHD in both SHS and WHI and showed that incorporating DNA

methylation data to predict CHD has substantial increase in prediction accuracy than only using traditional clinical features (risk factors). Finally, our AESurv model demonstrates the strength of learning low dimensional representations of high dimensional DNA methylation features. In the future, with the advancement of technology, even higher dimensional DNA methylation features may become available, and our model provides a new approach and complements the traditional survival analysis models in high-dimensional settings for more accurate CHD prediction.

Abbreviations

CHD = coronary heart disease; AESurv: deep learning autoencoder survival analysis; DeepSurv = Cox proportional hazards deep neural network model; CPH = Cox proportional hazard model; RSF = random survival forest; GBRTS = gradient boosted survival analysis; C-index = concordance index; AUROC = area under the receiver operating characteristic curve.

Key Points

- Developed a deep learning autoencoder survival analysis model (AESurv) that can incorporate high-dimensional DNA methylation and clinical data to learn lowdimensional participant representations towards coronary heart disease (CHD) prediction.
- The learned participant embeddings through AESurv can effectively reveal patterns of participants' CHD conditions.
- AESurv is able to accurately predict CHD in two different population cohorts: the Strong Heart Study (concordance index = 0.864 ± 0.009) and the Women's Health Initiative (concordance index = 0.752 ± 0.019) and achieved the best model performance compared with other machine learning models such as DeepSurv, random survival forest, and gradient boosting survival analysis models.
- The proposed AESurv model can be used to assist early detection of CHD based on DNA methylation and clinical features.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

The Strong Heart Study was supported by grants from the National Heart, Lung, and Blood Institute contracts 75N92019D00027, 75N92019D00028, 75N92019D00029, and 75N92019D00030; previous grants R01HL090863, R01HL109315, R01HL109301, R01HL109284, R01HL109282, and R01HL109319; and cooperative agreements U01HL41642, U01HL41652, U01HL41654, U01HL65520, and U01HL65521; and by National Institute of Environmental Health Sciences grants R01ES021367, R01ES025216, R01ES032638, P42ES033719, P30ES009089, and R35ES031688. We appreciate the participation of all Strong Heart Study participants and the support of the cohort staff.

The Women's Health Initiative program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005. A list of WHI Investigators is available at https://www.whi.org/doc/WHI-Investigator-Long-List.pdf. We appreciate the participation of all WHI participants and the support of the WHI Clinical Coordinating Center staff.

Data availability

The data were collected, analyzed, and reported under agreements made with the sovereign tribal nations that have partnered in this research, which precludes commonly accepted modes of data sharing. Requests to access the dataset from qualified researchers trained in human subject confidentiality protocols may be sent to the Strong Heart Study Coordinating Center at https://strongheartstudy.org/. Requests will be reviewed by tribal research partners before data may be released. This policy is consistent with the NIH Policy for Data Management and Sharing: Responsible Management and Sharing of American Indian/Alaska Native Participant Data [42]. Similarly, the procedures to request Women's Health Initiative data are detailed in the study website (https://www.whi.org/).

Author contributions

Y.S. and F.G. conceived the idea, designed research, developed AESurv model, developed code, analyzed data, and wrote the paper. A.N. and A.A.B. obtained funding and supervised the study. S.C., J.U., Y.Z. and A.F. obtained SHS funding. E.A.W. and A.A.B. obtained WHI funding. S.C. planned and conducted laboratory analyses. A.D. conducted DNA methylation and clinical data processing and coordination. Y.S., F.G., A.N., A.A.B., E.A.W., A.D., H.W., J.E.M., A.K., S.H., P.F.S., R.C., and M.K. contributed to analysis and interpretation of results. All authors have contributed to the manuscript preparation.

References

1. Wilson PW, D'Agostino RB, Levy D. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;**97**:1837–47. https://doi.org/10.1161/01.CIR.97.18.1837.
2. Murray CJ, Lopez AD. Mortality by cause for eight regions of the world: global burden of disease study. *The lancet* 1997;**349**:1269–76.
3. Stolpe S, Kowall B, Stang A. Decline of coronary heart disease mortality is strongly effected by changing patterns of underlying causes of death: an analysis of mortality data from 27 countries of the WHO European region 2000 and 2013. *Eur J Epidemiol* 2021;**36**:57–68. https://doi.org/10.1007/s10654-020-00699-0.
4. Mendis S, Puska P, Norrving B. *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization, Geneva, Switzerland, 2011.
5. Goff DC, Lloyd-Jones DM, Bennett G. *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. *J Am Coll Cardiol* 2014;**63**:2935–59. https://doi.org/10.1016/j.jacc.2013.11.005.
6. Navas-Acien A, Domingo-Relloso A, Subedi P. *et al.* Blood DNA methylation and incident coronary heart disease: evidence from the strong heart study. *JAMA Cardiol* 2021;**6**:1237–46. https://doi.org/10.1001/jamacardio.2021.2704.

7. Dogan MV, Grumbach IM, Michaelson JJ. *et al.* Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham heart study. *PloS One* 2018;**13**:e0190549. https://doi.org/10.1371/journal.pone.0190549.
8. Dogan MV, Beach SRH, Simons RL. *et al.* Blood-based biomarkers for predicting the risk for five-year incident coronary heart disease in the Framingham heart study via machine learning. *Genes* 2018;**9**:641. https://doi.org/10.3390/genes9120641.
9. Cugliari G, Benevenuta S, Guarrera S. *et al.* Improving the prediction of cardiovascular risk with machine-learning and DNA methylation data. In: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, Tuscany, Italy, 2019, 1–4.
10. Agha G, Mendelson MM, Ward-Caviness CK. *et al.* Blood leukocyte DNA methylation predicts risk of future myocardial infarction and coronary heart disease: a longitudinal study of 11 461 participants from population-based cohorts. *Circulation* 2019;**140**:645–57. https://doi.org/10.1161/CIRCULATIONAHA.118.039357.
11. Baccarelli A, Dolinoy DC, Walker CL. A precision environmental health approach to prevention of human disease. *Nat Commun* 2023;**14**:2449. https://doi.org/10.1038/s41467-023-37626-2.
12. Baccarelli A, Wright R, Bollati V. *et al.* Ischemic heart disease and stroke in relation to blood DNA methylation. *Epidemiology* 2010;**21**:819–28. https://doi.org/10.1097/EDE.0b013e3181f20457.
13. Byun HM, Colicino E, Trevisi L. *et al.* Effects of air pollution and blood mitochondrial DNA methylation on markers of heart rate variability. *J Am Heart Assoc* 2016;**5**:e003218. https://doi.org/10.1161/JAHA.116.003218.
14. Domingo-Relloso A, Makhani K, Riffo-Campos AL. *et al.* Arsenic exposure, blood DNA methylation, and cardiovascular disease. *Circ Res* 2022;**131**:e51–69. https://doi.org/10.1161/CIRCRESAHA.122.320991.
15. Domingo-Relloso A, Riffo-Campos AL, Haack K. *et al.* Cadmium, smoking, and human blood DNA methylation profiles in adults from the strong heart study. *Environ Health Perspect* 2020;**128**:067005. https://doi.org/10.1289/EHP6345.
16. Joehanes R, Just AC, Marioni RE. *et al.* Epigenetic signatures of cigarette smoking, circulation: cardiovascular. *Genetics* 2016;**9**:436–47.
17. Richard MA, Huan T, Ligthart S. *et al.* DNA methylation analysis identifies loci for blood pressure regulation. *The American Journal of Human Genetics* 2017;**101**:888–902. https://doi.org/10.1016/j.ajhg.2017.09.028.
18. Ligthart S, Marzi C, Aslibekyan S. *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol* 2016;**17**:1–15.
19. Dick KJ, Nelson CP, Tsaprouni L. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *The lancet* 2014;**383**:1990–8. https://doi.org/10.1016/S0140-6736(13)62674-4.
20. Turunen MP, Aavik E, Ylä-Herttuala S. Epigenetics and atherosclerosis. *Biochimica et Biophysica Acta (BBA)-General Subjects* 2009;**1790**:886–91.
21. Katzman JL, Shaham U, Cloninger A. *et al.* DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018;**18**:1–12. https://doi.org/10.1186/s12874-018-0482-1.
22. Illumina. *Comprehensive Coverage for Genome-Wide DNA Methylation Studies*. San Diego, CA, USA. https://www.illumina.com/techniques/microarrays/methylation-arrays.html (27 April 2022, date last accessed).
23. Bibikova M, Nazor KL, Altun G, Laurent LC Chapter 20 - epigenetics: DNA methylation. In: Loring J. F., Peterson S. E. eds). *Human Stem Cell Manual* (Second Edition). Boston: Academic Press, 2012, 325–36, https://doi.org/10.1016/B978-0-12-385473-5.00020-5.
24. Xia Y, Brewer A, Bell JT. DNA methylation signatures of incident coronary heart disease: findings from epigenome-wide association studies. *Clin Epigenetics* 2021;**13**:1–16. https://doi.org/10.1186/s13148-021-01175-6.
25. Li J, Zhu X, Yu K. *et al.* Genome-wide analysis of DNA methylation and acute coronary syndrome. *Circ Res* 2017;**120**:1754–67. https://doi.org/10.1161/CIRCRESAHA.116.310324.
26. Gao F, Zhang W, Baccarelli AA. *et al.* Predicting chemical ecotoxicity by learning latent space chemical representations. *Environ Int* 2022;**163**:107224. https://doi.org/10.1016/j.envint.2022.107224.
27. Assimes T, Tsao P, Absher D. *et al.* BA23–integrative genomics and risk of CHD and related phenotypes in the Women's health initiative. Women's Health Initiative. https://www.whi.org/study/BA23 (3 March 2024, date last accessed).
28. Vincent P, Larochelle H, Bengio Y *et al.* Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*. 2008, p. 1096–103.
29. Agarap AF. Deep learning using rectified linear units (relu). arXiv:1803.08375, 2018.
30. Yarotsky D. Error bounds for approximations with deep ReLU networks. *Neural Netw* 2017;**94**:103–14.
31. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. 2011, p. 315–23.
32. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
33. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008;**9**:2579–605.
34. Cox DR. Regression models and life-tables. *J R Stat Soc B Methodol* 1972;**34**:187–202.
35. Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical science* 1996;**11**:89–121.
36. Ishwaran H, Kogalur UB, Blackstone EH. *et al.* Random survival forests. *The annals of applied statistics* 2008;**2**:841–60. https://doi.org/10.1214/08-AOAS169.
37. Pölsterl S. Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J Mach Learn Res* 2020;**21**:1–6.
38. Harrell FE Jr, Lee KL, Califf RM. *et al.* Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;**3**:143–52. https://doi.org/10.1002/sim.4780030207.
39. Uno H, Cai T, Pencina MJ. *et al.* On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;**30**:1105–17. https://doi.org/10.1002/sim.4154.
40. Rakyan VK, Down TA, Balding DJ. *et al.* Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011;**12**:529–41. https://doi.org/10.1038/nrg3000.
41. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 2016;**8**:389–99. https://doi.org/10.2217/epi.15.114.
42. National Institutes of Health. *Supplemental Information to the NIH Policy for Data Management and Sharing: Responsible Management and Sharing of American Indian/Alaska Native Participant Data*. National Institute of Health, Bethesda, Maryland. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-214.html#:~:text=NIH%20recognizes%20that%20conducting%20biomedical,Tribes%5Biii%5D%20and%20communities (12 April 2023, date last accessed).