PLOS ONE

RESEARCH ARTICLE

Development and validation of risk prediction models for multiple cardiovascular diseases and Type 2 diabetes

Mehrdad Rezaee[1,2]*, **Igor Putrenko**[1], **Arsia Takeh**[1], **Andrea Ganna**[3,4,5], **Erik Ingelsson**[6,7,8]

1 Mynomx Inc., Palo Alto, CA, United States of America, **2** Cardiac and Vascular Care Inc., San Jose, CA, United States of America, **3** Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, United States of America, **4** Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, United States of America, **5** Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, United States of America, **6** Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States of America, **7** Stanford Cardiovascular Institute, Stanford, CA, United States of America, **8** Stanford Diabetes Research Center, Stanford, CA, United States of America

* mrezaee@mynomx.com

Citation: Rezaee M, Putrenko I, Takeh A, Ganna A, Ingelsson E (2020) Development and validation of risk prediction models for multiple cardiovascular diseases and Type 2 diabetes. PLoS ONE 15(7): e0235758. https://doi.org/10.1371/journal.pone.0235758

Editor: Helena Kuivaniemi, Stellenbosch University Faculty of Medicine and Health Sciences, SOUTH AFRICA

Received: December 26, 2019

Accepted: June 22, 2020

Published: July 29, 2020

Data Availability Statement: The data in this study is owned by a third party, the UK Biobank (www.ukbiobank.ac.uk) and legal constraints do not permit public sharing of the data. The UK Biobank, however, is open to all bona fide researchers anywhere in the world. Thus, the data used in this communication can be easily and directly accessed by applying through the UK Biobank Access Management System (www.ukbiobank.ac.uk/register-apply).

Abstract

Accurate risk assessment of an individuals' propensity to develop cardiovascular diseases (CVDs) is crucial for the prevention of these conditions. Numerous published risk prediction models used for CVD risk assessment are based on conventional risk factors and include only a limited number of biomarkers. The addition of novel biomarkers can boost the discriminative ability of risk prediction models for CVDs with different pathogenesis. The present study reports the development of risk prediction models for a range of heterogeneous CVDs, including coronary artery disease (CAD), stroke, deep vein thrombosis (DVT), and abdominal aortic aneurysm (AAA), as well as for Type 2 diabetes mellitus (DM2), a major CVD risk factor. In addition to conventional risk factors, the models incorporate various blood biomarkers and comorbidities to improve both individual and population stratification. An automatic variable selection approach was developed to generate the best set of explanatory variables for each model from the initial panel of risk factors. In total, up to 254,220 UK Biobank participants (ranging from 215,269 to 254,220 for different CVDs and DM2) were included in the analyses. The derived prediction models utilizing Cox proportional hazards regression achieved consistent discrimination performance (C-index) for all diseases: CAD, 0.794 (95% CI, 0.787–0.801); DM2, 0.909 (95% CI, 0.903–0.916); stroke, 0.778 (95% CI, 0.756–0.801); DVT, 0.743 (95% CI, 0.737–0.749); and AAA, 0.893 (95% CI, 0.874–0.912). When validated on various subpopulations, they demonstrated higher discrimination in healthier and middle-age individuals. In general, calibration of a five-year risk of developing the CVDs and DM2 demonstrated incremental overestimation of disease-related conditions amongst the highest decile of risk probabilities. In summary, the risk prediction models described were validated with high discrimination and good calibration for several CVDs and DM2. These models incorporate multiple shared predictor variables and may be integrated into a single platform to enhance clinical stratification to impact health outcomes.

Funding: Mynomx, Inc. provided funding for this study in the form of salaries for authors AT and IP, consultancy fees to AG, and as an unrestricted research grant to Stanford University (led by EI). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'Author contributions' section.

Competing interests: The authors have read the journal's policy, and the authors have the following competing interests to declare: AT and IP are employees of Mynomx, Inc. AG received consultancy fees from Mynomx, Inc. MR is on the board of Mynomx, Inc., but did not receive any financial compensation for participation. This does not alter our adherence to PLOS ONE policies on sharing data and materials. There are no patents, products in development or marketed products associated with this research to declare.

Introduction

Cardiovascular diseases (CVDs) include a range of chronic diseases that impair cardiac and vascular function, which continues to be the leading cause of death in the United States (US). It is projected that over 45% of the US population will suffer from one or more CVDs by 2035 [1]. The healthcare cost associated with CVDs represents one of the greatest global economic burdens [2].

As with any chronic condition, appropriate prevention and selective treatment of CVDs are the most effective approaches to reduce their clinical and financial impact. Accurate risk assessment of an individual's propensity to develop CVDs is essential for personalized health care and primary prevention of these conditions. An increasing number of novel biomarkers have been linked to CVD risk [3–14], implying their critical role in precise risk assessment for heterogeneous CVDs. Current established functions for CVD risk stratification are either based only on conventional risk factors or include a limited number of biomarkers [15–18]. Furthermore, the contribution of various biomarkers to the risk of CVDs with different pathogenesis is poorly understood.

In this study, we sought to improve CVD risk stratification through the addition of multiple blood biomarkers in CVD risk prediction modeling. We report the development and validation of risk prediction models for a range of heterogeneous CVDs with different pathogenesis, including coronary artery disease (CAD), stroke, deep venous thrombosis (DVT), and abdominal aortic aneurysm (AAA), as well as for Type 2 diabetes mellitus (DM2), a prime CVD risk factor [19]. The aforementioned diseases together are broadly defined in the present study as cardiometabolic diseases (CMDs). The prediction models were derived using a large population (UK Biobank [20]) analysis with a median longitudinal follow-up of 6.1 years and incorporated a distinct combination of conventional risk factors, blood biomarkers, and comorbidities produced by uncurated variable selection.

Materials and methods

Inclusion/exclusion criteria and outcome definition

Baseline data for 502,616 UK Biobank (UKBB) participants collected at assessment centers were used to derive the prediction models. Overall, 95% of the UKBB participants were self-described as white, with women comprising 54.4% of the participant population. The UKBB data was subsequently linked to hospital episode statistics (HES) data from hospitals in England, Scotland, and Wales (**Fig 1**). Outcomes for coronary artery disease (CAD), Type 2 diabetes mellitus (DM2), stroke, deep venous thrombosis (DVT), and abdominal aortic aneurysm (AAA) were determined according to documentation using the following International Classification of Diseases (ICDs) for each of the diseases:

1. International Classification of Diseases edition 10 (ICD-10) codes for all CMD outcomes in the HES data. The following ICDs were used: I20–I25 and T82 codes were used for CAD; E11-E14 codes for DM2; G46.3, G46.4, I63, I66, I67, and I693 codes for stroke; I82, O22.3, R09.89, and Z86.7 codes for DVT; and I71 and I79.0 codes for AAA.
2. Self-reporting for CAD, DM2, and DVT.
3. Self-reported medications for CAD (Nitrolingual, Tildiem) and DM2 (Rosiglitazone, Pioglitazone, Metformin, Isosorbide mononitrate, Insulin products, Glucophage, Glimepiride, Gliclazide).

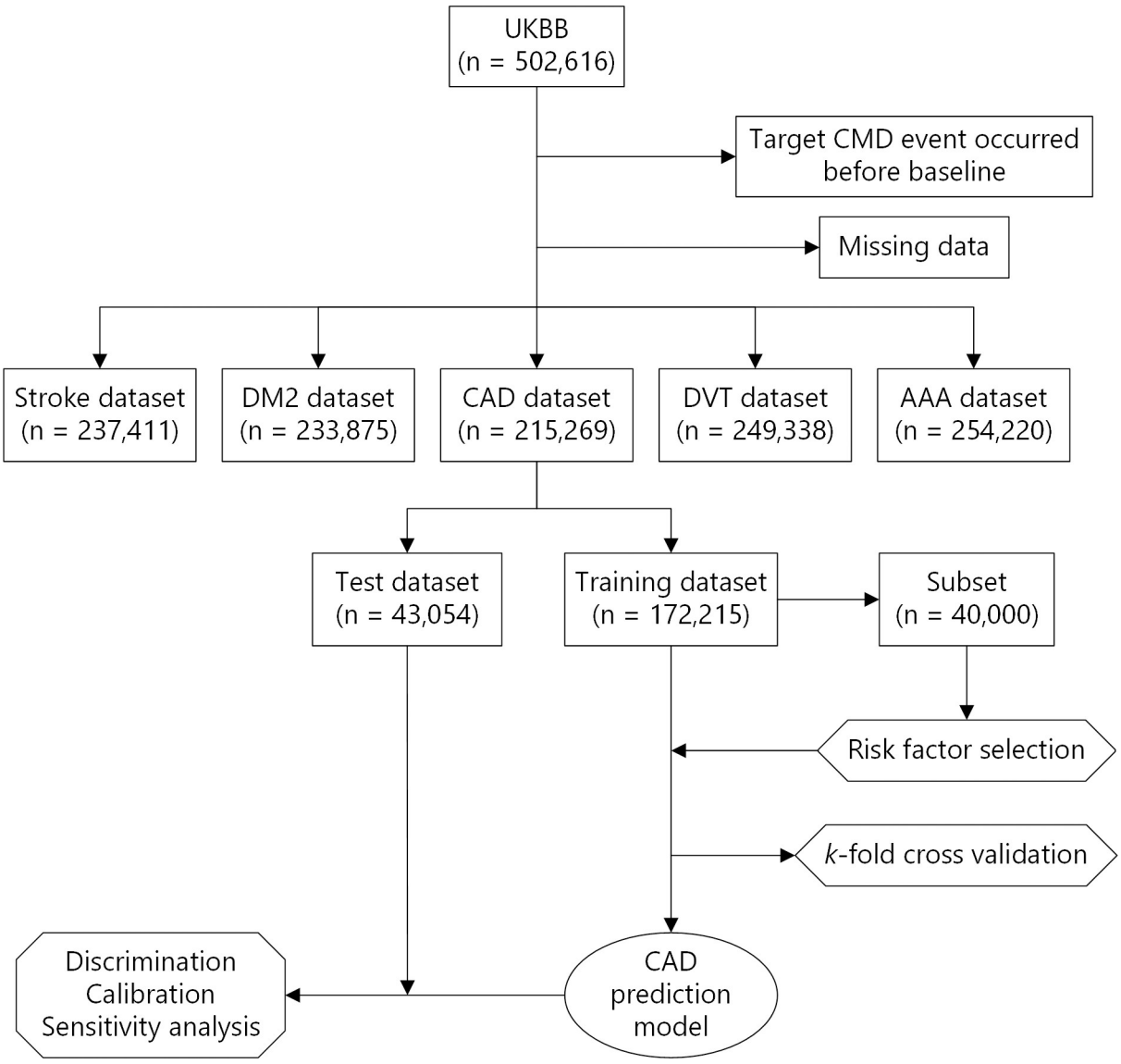

Fig 1. Flow chart demonstrating the exclusion criteria, and the development process of the CAD model, and subsequent validation. The process was the same for the other CMDs.

https://doi.org/10.1371/journal.pone.0235758.g001

The age and date of a CMD event were determined based on primary or secondary ICD-10 codes in the HES data corresponding to the earliest hospitalization records. Individuals with more than one CMD diagnosis during a given admission were included in the study samples for corresponding CMD. The date of inclusion in the UKBB was defined as the participant's baseline and was used as the starting point for time-to-event calculations. Participants with a target CMD event before baseline (identified by ICD-10 codes, self-reports, or medication) were excluded from the study sample for the corresponding CMD modeling (Fig 1). However, participants with prior non-target CMD event(s) (potential comorbidity) were not excluded. The above exclusions resulted in five CMD-specific datasets with samples sizes ranging from approximately 466,000 to 481,000. Incident cases were defined as CMD-positive cases per ICD-10 codes and had the date of the event recorded on the HES data after the baseline. Self-reported diagnoses and medications were only used to identify prevalent cases since this

information is only available at baseline. Further exclusion of cases due to missing data produced a final five study populations used to develop the prediction models; these had sample sizes of 215,269 (CAD), 233,875 (DM2), 237,411 (stroke), 249,338 (DVT), and 254,220 (AAA) (Fig 1). The exit date was determined as the occurrence of either date of death, end of follow-up (February 29, 2016), or a CMD event.

Risk factors for predictive modeling

Conventional CMD risk factors in the prediction models were selected according to frequency of documentation. Accordingly, the variables selected were missing from less than 80,000 individuals. The list of these risk factors included: age, gender, body mass index (BMI), systolic and diastolic blood pressure (SBP and DBP), physical activity, current and past smoking history, and family history. Physical activity was assessed as the metabolic equivalent of task (MET) and calculated in hours/week according to the "Guidelines for Data Processing and Analysis of the International Physical Activity Questionnaire (IPAQ)" [21]. Family history included mother or father for DM2 and stroke, and mother, father, or siblings for CAD. Binary variables describing these combinations for family history were used in predictive modeling.

Additionally, 22 blood biomarkers, including three blood count tests and 19 biochemical markers, were considered as risk factors. We further considered novel risk factors: self-reported sleep apnea, congestive heart failure, arrhythmia, heart valve problem, irritable bowel syndrome, and hyperthyroidism. The arrhythmia category included atrial fibrillation, atrial flutter, Wolff-Parkinson-White (WPW) syndrome, irregular heartbeat, sick sinus syndrome, and supraventricular tachycardia. The heart valve deficiency category included mitral valve prolapse, mitral stenosis, mitral regurgitation/incompetence, aortic valve disease, aortic stenosis, and aortic regurgitation/incompetence.

Data preparation and variable selection for predictive modeling

Python 3.6.6 for Windows x64 was used for the preparation of datasets for each CMD according to the approaches described above. The datasets were further split into 80% training and 20% testing sets using the pseudorandom number generator algorithm with a constant initial seed value of 42 (Fig 1). Training sets were used for model fitting and assessment and variable selection. Testing sets were solely used for assessing models' discrimination performance and calibration as well as for sensitivity analysis.

The Recursive Feature Elimination (RFE) method (scikit-learn 0.20.0 Python library) was used to automatically construct the best set of predictor variables for each CMD model from the initially available panel of candidate risk factors. Multiple random forest binary classification models predicting the occurrence of a CMD event by the end of the follow-up period were constructed based on subsets of variables of decreasing size, and the models' performance was compared [22]. Considering low CMD incidence rate, we used the Balanced Random Forest algorithm (imbalanced-learn 0.4.2 Python library) downsampling majority class to balance it with minority class in a bootstrap sample in each decision tree [23]. The accuracy of classification determined as the fraction of correct predictions was used for the models' performance evaluation.

The RFE method was combined with a stratified two-fold cross-validation using the following procedure: 1) 40,000 samples were randomly selected from a training set and split into two equal sized subsets with preserved ratio of positive to negative CMD cases; 2) variables were recursively removed one-by-one, and a model using the remaining variables was fitted to one subset, and its accuracy was evaluated on another subset; 3) this process was run once on each subset, and average accuracy was calculated and used for ranking and removing the weakest

variables and selecting the best subset of variables. The RFE variable selection was applied to each CMD separately. The gender variable was forced into all CMD-specific sets of explanatory variables.

Additional variable selection based on a variance inflation factor (VIF) detecting correlation between variables was conducted for the DM2 model prior to RFECV to achieve better calibration. VIF_i for each variable was calculated using the following formula:

$$VIF_i = \frac{1}{1 - R_i^2},$$

where R_i^2 is the coefficient of determination for each variable. R_i^2 was calculated by regressing the variable against every other variable using ordinary least squares regression (statsmodels 0.9.0 Python library). Variables with the lowest VIF among all variables with VIF higher than 2 were removed one-by-one until all variables had VIF lower than 2. The VIF variable selection did not improve the calibration when applied to the rest of the CMD models.

Prediction models and performance metrics

Linear Cox Proportional Hazard (PH) models were developed using lifelines 0.13.0 Python library. Continuous variables were scaled to a range between 0 and 1 to allow for a comparison of the magnitudes of regression coefficients. The discriminative ability of the risk prediction models was assessed by Harrell's concordance index (C-index) [24–26], which was calculated during the validation and datasets testing as the proportion of all comparable pairs where the predictions and outcomes were concordant. Case pairs were comparable if at least one of them was CMD positive. If the prognostic index was larger for the case with a lower survival time, the prediction of that pair was counted as concordant. If predictions were identical for a pair, 0.5 was added to the count of concordance. A pair was not comparable if an event occurred for both at the same time or an event occurred for one, but the time of censoring was smaller than the time of event of the first one.

k-fold cross-validation was used to assess for overfitting leading to model optimism and to adjust estimates of discriminative ability for this optimism [27]. A training set was randomly partitioned into five complementary equal sized subsets. Of the five subsamples, a single subsample was retained as the validation set for testing the discriminative ability of a model, and the remaining four subsets were used as the training set. This process was repeated five times, with each of the five subsets used exactly once as the validation set. The resulting C-indexes were averaged to produce a single, overall optimism-corrected estimate of the C-index with a 95% confidence interval (CI) and standard deviation (SD).

Calibration of Cox PH models was evaluated by the Hosmer-Lemeshow goodness-of-fit test [28]. The Hosmer-Lemeshow test statistic was calculated using the following formula:

$$H = \sum_{g=1}^{G} \frac{(O_{1g} - E_{1g})^2}{N_g \pi_g (1 - \pi_g)},$$

where O_{1g} is observed CMD events, E_{1g} is expected CMD events, N_g is total observations, and π_g is predicted probability for the g^{th} risk decile group, and G is the number of groups. The testing set was divided into decile groups based on the predicted probability of CMD events for a time horizon of five years. Then, the number of observed CMD events and the sum of the predicted probabilities of CMD events (interpreted as the number of expected CMD events) were calculated in each decile group. The computed Hosmer-Lemeshow statistic was compared to a *chi*-squared distribution with eight (G-2) degrees of freedom to calculate the *P*-

value. Calibration was visualized using a calibration plot using the predicted risk probabilities plotted against the observed risks for each decile group.

Subpopulation sensitivity analysis

The performance and sensitivity of prediction models were assessed in four subgroups of individuals with different age and CMD status. Multiple testing sets were created by applying age and disease filters to the general testing datasets. These subpopulations included (1) "healthy" participants without any of four non-target CMD at the baseline; (2) "unhealthy" participants with one or more pre-existing non-target CMD at the baseline; (3) participants with only one non-target CMD (CAD, DM2, or DVT); and (4) participants in the age categories <45, 45–55, 55–65, and 65–75 years.

Results

The mean (SD) age at baseline was 56 (8.0) years across all CMD study samples, with women accounting for 54.0% to 55.6% of the participants in these samples. Gender-specific demographics, physiological and lifestyle characteristics, the number of CMD incident events (Table 1), as well as biochemical and clinical characteristics (S1 Table), showed no significant variation across different CMD study samples. Assessment of the participant's follow-up (median 6.1 years) reports, the incidence rates for CAD (3.32%), DM2 (2.65%), stroke (0.66%), DVT (2.44%), and AAA (0.17%) were observed in the corresponding study sample.

The discriminative ability of all Cox PH CMD models estimated by five-fold cross-validation varied between the diseases with the highest and lowest C-indexes for DM2 and DVT, respectively. The optimism-corrected estimate of discrimination C-statistic was 0.794 (CI, 0.787–0.801, SD = 0.0050) for CAD, 0.909 (CI, 0.903–0.916, SD = 0.0046) for DM2, 0.778 (CI, 0.756–0.801, SD = 0.0162) for stroke, 0.743 (CI, 0.737–0.749, SD = 0.0044) for DVT, 0.893 (CI, 0.874–0.912, SD = 0.0137) for AAA. A low standard deviation of C-statistic values for the CAD, DM2, and DVT models implied their high reproducibility and good generalization to unknown data from the same population and a low degree of overoptimism (Table 2). The models for stroke and AAA, the diseases with lower numbers of incident events, demonstrated a lower reproducibility compared to the other models. Performance assessment in testing sets of the general population (Table 2) demonstrated that C-indexes for the CAD and stroke models were within the above 95% CIs. C-indexes for the DM2, DVT, and AAA models were outside of the CIs by 0.003, 0.001, and 0.005, respectively, consistent with the low degree of overoptimism estimated by the cross-validation method.

The number of predictors in the best sets generated for different models varied from nine predictors for AAA to 40 for CAD (S2 Table). Among multiple biomarker predictors shared across different models, cystatin C and red blood cell distribution width were common risk factors for all four CVDs, but not for DM2. Comparison of the values of normalized regression coefficients (S2 Table) demonstrated that cystatin C was the most crucial risk factor for stroke, DVT, and AAA, and was superseded by glycated hemoglobin only in the CAD model. Glycated hemoglobin also was the most important risk factor for DM2 and was shared among stroke, DVT, and DM2; however, it was not a statistically significant variable for AAA. Overall, the CAD and stroke models shared the largest number of predictors among all diseases.

Broad range applicability and performance consistency for the developed risk prediction models for each disease was further determined by assessing the discriminative ability across subpopulations using sensitivity analysis (Table 2). This analysis demonstrated lower prediction in the "unhealthy" compared to the "healthy population," as defined in the material and

Table 1. Baseline characteristics of the study samples for each dataset used to derive CMD-specific prediction models.

	CAD		DM2		Stroke		DVT		AAA	
	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men
	n = 119,756	n = 95,513	n = 127,929	n = 105,946	n = 129,255	n = 108,156	n = 135,640	n = 115,075	n = 137,352	n = 116,868
Age, mean (SD), y	55 (8)	56 (8)	55 (8)	56 (8)	56 (8)	56 (8)	56 (8)	56 (8)	56 (8)	56 (8)
Body mass index, mean (SD)	26.4 (4.8)	27.3 (4.0)	26.4 (4.8)	27.3 (3.9)	26.5 (4.8)	27.4 (4.0)	26.5 (4.8)	27.4 (4.0)	26.5 (4.8)	27.5 (4.0)
Systolic blood pressure, mean (SD), mm Hg	133 (18)	139 (17)	133 (18.4)	139 (17)	133 (18)	139 (17)	133 (18)	140 (17)	133 (18)	140 (17)
Diastolic blood pressure, mean (SD), mm Hg	80 (10)	84 (10)	80 (10)	84 (10)	80(10)	83 (10)	80 (10)	83 (10)	80 (10)	83 (10)
Forced expiratory volume, mean (SD), %	94.1 (17.7)	93.1 (18.2)	94.1 (17.7)	93.1 (18.3)	94.0 (17.7)	92.8 (18.4)	93.9 (17.8)	92.6 (18.4)	93.9 (17.8)	92.5 (18.5)
Physical activity, mean (SD), MET x hours/week	48.8 (56.8)	59.6 (77.6)	48.7 (56.7)	59.3 (76.9)	48.7 (56.7)	59.0 (76.7)	48.9 (56.3)	59.6 (77.7)	48.9 (57.3)	59.4 (77.6)
Current smoking, No. (%)	9599 (8.02)	10733 (11.2)	10177 (7.96)	11836 (11.2)	10259 (7.94)	12030 (11.1)	11100 (8.18)	13237 (11.5)	11285 (8.22)	13484 (11.5)
Past smoking, No. (%)	65186 (54.4)	60297 (63.1)	69565 (54.4)	67026 (63.3)	70287 (54.4)	68581 (63.4)	74058 (54.6)	73217 (63.6)	75046 (54.6)	74517 (63.8)
Family history, No. (%)	58743 (49.1)	42320 (44.3)	10910 (8.53)	8304 (7.84)	33770 (26.1)	26599 (24.6)	N/A	N/A	N/A	N/A
CMD incident events, No. (%)	2476 (2.07)	4677 (4.90)	2387 (1.87)	3789 (3.58)	617 (0.48)	956 (0.88)	3350 (2.47)	4045 (3.52)	54 (0.039)	378 (0.32)

https://doi.org/10.1371/journal.pone.0235758.t001

methods. The performance of the models was highest in middle age (45–65 years), but it significantly dropped in the subpopulations with pre-existing CMD.

To evaluate the calibration properties of the prediction models amongst the general population, the five-year absolute risk for each CMD event was calculated. Hosmer-Lemeshow test statistic produced *chi*-squares values (*P*-value) of 33.8 (<0.0001), 77.8 (<0.0001), 12.8 (0.12), 45.3 (<0.0001), and 12.3 (0.14) for the CAD, DM2, stroke, DVT, and AAA models, respectively. Calibration plots (Fig 2) demonstrated consistent overall calibration, but overestimation of CMD risk in the highest decile of risk probabilities for all except the DM2 model. DM2 risk was slightly overestimated in the lowest deciles and minimally underestimated in the highest risk decile. As expected, with a low number of events, the prediction model for AAA was poorly calibrated.

Table 2. Discriminative ability of CMD risk prediction models among different subpopulations.

Test subpopulation	CAD	DM2	Stroke	DVT	AAA
General	0.789 (3.32)	0.9 (2.65)	0.776 (0.66)	0.750 (2.95)	0.869 (0.17)
Healthy + target CMD	0.788 (3.16)	0.903 (2.5)	0.76 (0.57)	0.738 (2.60)	0.874 (0.14)
Unhealthy + target CMD	0.643 (10.8)	0.752 (6.78)	0.65 (2.70)	0.575 (12.4)	0.710 (0.76)
CAD	N/A	0.727 (8.01)	0.663 (2.80)	0.582 (12.3)	0.745 (1.13)
DM2	0.655 (10.9)	N/A	0.646 (3.27)	0.594 (11.3)	N/A
DVT	0.644 (10.9)	0.763 (5.42)	0.633 (3.24)	N/A	N/A
Age < 45	0.741 (0.83)	0.898 (0.8)	0.524 (0.14)	0.645 (0.91)	N/A
Age 45–55	0.742 (1.57)	0.889 (1.51)	0.694 (0.25)	0.677 (1.42)	0.854 (0.03)
Age 55–65	0.734 (3.96)	0.897 (3.03)	0.696 (0.70)	0.705 (3.37)	0.778 (0.17)
Age 65–75	0.724 (7.29)	0.853 (5.19)	0.670 (1.71)	0.656 (6.15)	0.808 (0.51)

Mean C-indexes for CMD models and percent of participants (in parenthesis) that encountered target CMD incident events during the follow-up period.

https://doi.org/10.1371/journal.pone.0235758.t002

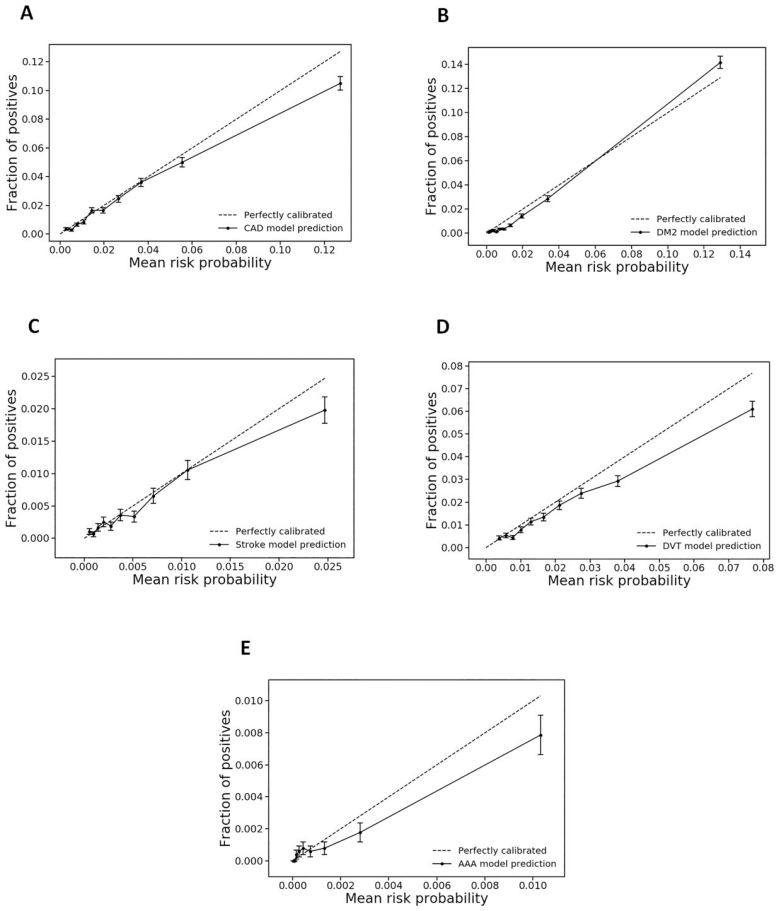

Fig 2. Calibration plots for CMD risk prediction models. Five-year absolute risks for CAD (A), DM2 (B), stroke (C), DVT (D), and AAA (E) were split into deciles, and mean risk probability for each decile was plotted versus the portion of positive CMD cases in the decile for a time horizon of five years.

https://doi.org/10.1371/journal.pone.0235758.g002

The mean risk probability versus mean survival time for each decile was plotted (Fig 3) to evaluate the discriminative ability of individual risk models. The mean risk probabilities for all CMDs exponentially decreased with increased survival time. A steeper exponential decay was observed for the DM2 and AAA models, which demonstrated the highest discriminative ability for these two diseases.

Discussion

In the present study, the development and validation of risk prediction models for a range of heterogeneous diseases with different pathogenesis, including four CVDs and DM2, were presented. Extensive population data from UK Biobank was applied to produce model-specific training, validation and testing sets. The discriminative ability of individual models for each of the chronic diseases was first examined in the general population. Subsequently, the impact of comorbidities amongst various age groups was determined. The discrimination performances were high in the overall UK Biobank population and remained moderate to high in the sub-population analysis. Calibration of the five-years survival outcome demonstrated incremental overestimation of disease-related conditions amongst the highest decile of risk probabilities. Internal validation of the developed models demonstrated good reproducibility and a low degree of overoptimism.

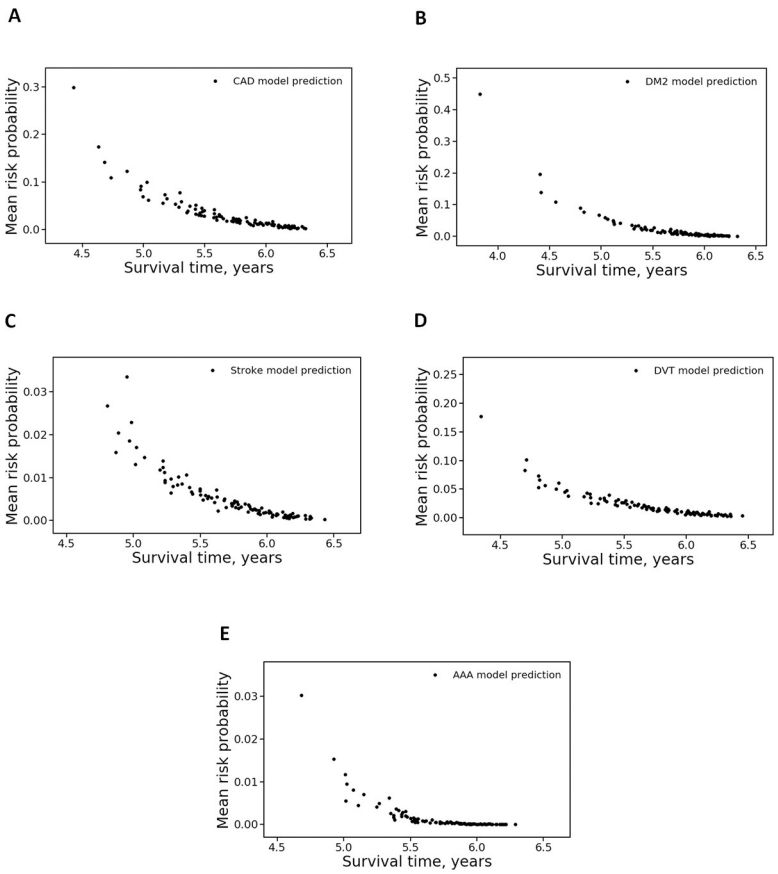

Fig 3. CMD risk prediction models demonstrate the relation between risk level and survival. Risk probabilities for five diseases, CAD (A), DM2 (B), stroke (C), DVT (D), and AAA (E), were split into deciles, and mean risk probability for each decile was plotted versus mean survival time in the decile.

https://doi.org/10.1371/journal.pone.0235758.g003

In addition to conventional risk factors, the models described in this study incorporated multiple blood biomarkers and comorbidities. Contemporary risk factors such as biomarkers, polygenic risk scores [13, 29–31], and certain metabolomic patterns [32, 33] have been proposed to augment the total risk assessment. As demonstrated previously, depending on the population, up to 50% of patients with CVDs may lack conventional risk factors. However, these conventional risk factors can also fail to identify between 15–50% of those at risk of developing cardiovascular disease [34–38]. Understanding and differentiating between clinical statuses due to acquired risk factors versus genetic predispositions can significantly impact the approaches to risk factors modification, which can change the course of disease progression. Future work will focus on exploring the value of polygenic risk scores to improve the performances of the models described in this report.

The predictive modeling presented also demonstrated that many of the risk factors are shared across various CVDs and DM2, implying complex pathophysiological links. Positive associations reported previously between CAD, stroke, AAA, and DVT, with both cystatin C and red blood cell distribution width [3–8, 10–12, 14], as well as similar, atherosclerosis-based pathogenesis of CAD and stroke also support our observations.

An automatic approach for variable selection developed in this study allowed us to produce disease-specific sets of explanatory variables and to include novel biomarkers that were not previously used in risk stratification for CVDs or DM2. Prediction models for heterogeneous

diseases constructed using the same general panel of candidate predictors had a good predictive performance and reproducibility. This is the novelty of our work as compared to previously published risk prediction models for composite CVD (myocardial infarction, angina, coronary heart disease, stroke, and transient ischemic attack) [39], DVT [40], AAA [41] and DM2 [42] that incorporated conventional risk factors selected by a labor-intensive curated process. When applied to large-scale disease-agnostic datasets with a large number of potential predictors derived from electronic health records (EHR), domain knowledge-based variable selection can discard important information. Increasing use of comprehensive EHR for more accurate risk stratification and prediction of patient outcomes [43] further underlines the importance of application of automatic approaches to variable selection.

Limitations of this study

Given the predominantly white UK Biobank population and the fact that both training and testing datasets were produced from the same population, the developed prediction models are unlikely to be generalizable to other populations, and their transportability requires further assessment in external validation studies. Low transportability is a common limitation of prediction models, including established CVD risk algorithms. It was reported that neither the Framingham (derived from a US cohort [15]) nor ASSIGN (derived from the Scottish Heart Health Extended Cohort [44]) algorithms were well calibrated for the UK population, with both scores tending to over-predict risk [45]. These algorithms also had a decreased discrimination performance in comparison to the QRISK algorithm derived from a large primary care database in England and Wales [45].

Comparison of risk factor associations in UK Biobank against representative, general population a recent study by Batty et al. [46] demonstrated that associations between the risk factors and health outcomes are generalizable. Accordingly, recalibration of risk scores can be performed using these associations and an updated baseline risk for a specific population. It should also be emphasized that even a well-calibrated risk algorithm does not automatically translate to improved patient outcomes. Substantial work is required to make CVD risk stratification a practical and effective clinical tool.

Future directions

Considering computational limitations of non-linear survival models [47], binary classification models utilizing deep learning algorithms can be adapted in the future to determine the probability of CMD events at certain time horizons. The availability of relatively large healthcare datasets with thousands of potential predictors further supports the application of deep learning in CVD risk assessment. In addition, incorporation of genomic and other omics data may further improve the predictive functionality provided by the developed models.

Supporting information

S1 Table. Baseline biochemical and clinical characteristics of the study samples for each dataset used to derive CMD-specific prediction models.
(XLSX)

S2 Table. Normalized Cox PH model regression coefficients for five CMDs. Regression coefficients (coef) and corresponding standard errors (se), *P*-values, lower and upper 95% CI are presented.
(XLSX)

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 24626 to Mynomx, Inc.

Author Contributions

Conceptualization: Mehrdad Rezaee, Andrea Ganna.

Data curation: Igor Putrenko, Arsia Takeh.

Formal analysis: Igor Putrenko, Arsia Takeh, Andrea Ganna, Erik Ingelsson.

Funding acquisition: Mehrdad Rezaee.

Investigation: Arsia Takeh, Erik Ingelsson.

Methodology: Igor Putrenko, Arsia Takeh, Andrea Ganna, Erik Ingelsson.

Project administration: Mehrdad Rezaee.

Software: Igor Putrenko, Arsia Takeh.

Supervision: Mehrdad Rezaee, Erik Ingelsson.

Validation: Igor Putrenko, Arsia Takeh, Andrea Ganna.

Writing – original draft: Mehrdad Rezaee, Igor Putrenko, Arsia Takeh.

Writing – review & editing: Mehrdad Rezaee, Igor Putrenko, Arsia Takeh, Andrea Ganna, Erik Ingelsson.

References

1. Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, et al. American Heart Association Council on epidemiology and prevention statistics committee and stroke statistics subcommittee. Heart disease and stroke statistics-2018 update: a report from the American Heart Association Circulation. 2018; 137(12):e67–e492. https://doi.org/10.1161/CIR.0000000000000558 PMID: 29386200
2. Muka T, Imo D, Jaspers L, Colpani V, Chaker L, van der Lee SJ, et al. The global impact of non-communicable diseases on healthcare spending and national income: a systematic review. Eur J Epidemiol. 2015; 30(4):251–77. https://doi.org/10.1007/s10654-014-9984-2 PMID: 25595318
3. Balistreri CR, Pisano C, Bertoldo F, Massoud R, Dolci S, Ruvolo G. Red Blood Cell Distribution Width, Vascular Aging Biomarkers, and Endothelial Progenitor Cells for Predicting Vascular Aging and Diagnosing/Prognosing Age-Related Degenerative Arterial Diseases. Rejuvenation Res. 2019; 22(5):399–408. https://doi.org/10.1089/rej.2018.2144 PMID: 30572793
4. Bell EJ, Selvin E, Lutsey PL, Nambi V, Cushman M, Folsom AR. Glycemia (hemoglobin A1c) and incident venous thromboembolism in the Atherosclerosis Risk in Communities cohort study. Vasc Med. 2013; 18(5):245–50. https://doi.org/10.1177/1358863X13506764 PMID: 24165467
5. Brodin EE, Braekkan SK, Vik A, Brox J, Hansen JB. Cystatin C is associated with risk of venous thromboembolism in subjects with normal kidney function—the Tromso study. Haematologica. 2012; 97 (7):1008–13. https://doi.org/10.3324/haematol.2011.057653 PMID: 22315498
6. Cay N, Unal O, Kartal MG, Ozdemir M, Tola M. Increased level of red blood cell distribution width is associated with deep venous thrombosis. Blood coagulation & fibrinolysis. 2013; 24(7):727–31.
7. Gregson J, Kaptoge S, Bolton T, Pennells L, Willeit P, Burgess S, et al. Cardiovascular Risk Factors Associated With Venous Thromboembolism. JAMA Cardiol. 2019; 4(2):163–73. https://doi.org/10.1001/jamacardio.2018.4537 PMID: 30649175
8. Khaw K-T, Wareham N. Glycated hemoglobin as a marker of cardiovascular risk. Current opinion in lipidology. 2006; 17(6):637–43. https://doi.org/10.1097/MOL.0b013e3280106b95 PMID: 17095908
9. Kristensen KL, Dahl M, Rasmussen LM, Lindholt JS. Glycated Hemoglobin Is Associated With the Growth Rate of Abdominal Aortic Aneurysms: A Substudy From the VIVA (Viborg Vascular) Randomized Screening Trial. Arterioscler Thromb Vasc Biol. 2017; 37(4):730–6. https://doi.org/10.1161/ATVBAHA.116.308874 PMID: 28183702

10. Li N, Zhou H, Tang Q. Red Blood Cell Distribution Width: A Novel Predictive Indicator for Cardiovascular and Cerebrovascular Diseases. Dis Markers. 2017; 2017:7089493. https://doi.org/10.1155/2017/7089493 PMID: 29038615

11. Lv BJ, Lindholt JS, Cheng X, Wang J, Shi GP. Plasma cathepsin S and cystatin C levels and risk of abdominal aortic aneurysm: a randomized population-based study. PLoS One. 2012; 7(7):e41813. https://doi.org/10.1371/journal.pone.0041813 PMID: 22844527

12. Mitsios JP, Ekinci EI, Mitsios GP, Churilov L, Thijs V. Relationship Between Glycated Hemoglobin and Stroke Risk: A Systematic Review and Meta-Analysis. J Am Heart Assoc. 2018; 7(11).

13. Wang J, Tan GJ, Han LN, Bai YY, He M, Liu HB. Novel biomarkers for cardiovascular risk prediction. J Geriatr Cardiol. 2017; 14(2):135–50. https://doi.org/10.11909/j.issn.1671-5411.2017.02.008 PMID: 28491088

14. Wattanakit K, Lutsey PL, Bell EJ, Gornik H, Cushman M, Heckbert SR, et al. Association between cardiovascular disease risk factors and occurrence of venous thromboembolism. A time-dependent analysis. Thromb Haemost. 2012; 108(3):508–15. https://doi.org/10.1160/TH11-10-0726 PMID: 22782466

15. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991; 121(1 Pt 2):293–8.

16. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003; 24(11):987–1003. https://doi.org/10.1016/s0195-668x(03)00114-3 PMID: 12788299

17. D'Agostino RB Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008; 117(6):743–53. https://doi.org/10.1161/CIRCULATIONAHA.107.699579 PMID: 18212285

18. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ. 2010; 341:c6624. https://doi.org/10.1136/bmj.c6624 PMID: 21148212

19. Dokken BB. The pathophysiology of cardiovascular disease and diabetes: beyond blood pressure and lipids. Diabetes Spectrum. 2008; 21(3):160–5.

20. Palmer LJ. UK Biobank: bank on it. Lancet. 2007; 369(9578):1980–2. https://doi.org/10.1016/S0140-6736(07)60924-6 PMID: 17574079

21. Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc. 2003; 35(8):1381–95. https://doi.org/10.1249/01.MSS.0000078924.61453.FB PMID: 12900694

22. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002; 46(1–3):389–422.

23. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. University of California, Berkeley. 2004; 110(1–12):24.

24. Harrell FE Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982; 247(18):2543–6. PMID: 7069920

25. Harrell FE Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med. 1984; 3(2):143–52. https://doi.org/10.1002/sim.4780030207 PMID: 6463451

26. Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15(4):361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4 PMID: 8668867

27. Smith GC, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. Am J Epidemiol. 2014; 180(3):318–24. https://doi.org/10.1093/aje/kwu140 PMID: 24966219

28. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997; 16(9):965–80. https://doi.org/10.1002/(sici)1097-0258(19970515)16:9<965::aid-sim509>3.0.co;2-o PMID: 9160492

29. Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med. 2008; 358(12):1240–9. https://doi.org/10.1056/NEJMoa0706728 PMID: 18354102

30. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018; 50(9):1219–24. https://doi.org/10.1038/s41588-018-0183-z PMID: 30104762

31. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. N Engl J Med. 2016; 375(24):2349–58. https://doi.org/10.1056/NEJMoa1605086 PMID: 27959714

32. Ganna A, Salihovic S, Sundstrom J, Broeckling CD, Hedman AK, Magnusson PK, et al. Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. PLoS Genet. 2014; 10(12):e1004801. https://doi.org/10.1371/journal.pgen.1004801 PMID: 25502724

33. Ruiz-Canela M, Hruby A, Clish CB, Liang L, Martinez-Gonzalez MA, Hu FB. Comprehensive Metabolomic Profiling and Incident Cardiovascular Disease: A Systematic Review. J Am Heart Assoc. 2017; 6 (10).

34. Futterman LG, Lemberg L. Fifty percent of patients with coronary artery disease do not have any of the conventional risk factors. Am J Crit Care. 1998; 7(3):240–4. PMID: 9579251

35. Hennekens CH. Increasing burden of cardiovascular disease: current knowledge and future directions for research on risk factors. Circulation. 1998; 97(11):1095–102. https://doi.org/10.1161/01.cir.97.11.1095 PMID: 9531257

36. Khot UN, Khot MB, Bajzer CT, Sapp SK, Ohman EM, Brener SJ, et al. Prevalence of conventional risk factors in patients with coronary heart disease. JAMA. 2003; 290(7):898–904. https://doi.org/10.1001/jama.290.7.898 PMID: 12928466

37. Lefkowitz RJ, Willerson JT. Prospects for cardiovascular research. JAMA. 2001; 285(5):581–7. https://doi.org/10.1001/jama.285.5.581 PMID: 11176863

38. McKechnie RS, Rubenfire M. The role of inflammation and infection in coronary artery disease: a clinical perspective. ACC Current Journal Review. 2002; 1(11):32–4.

39. van Staa TP, Gulliford M, Ng ES, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? PLoS One. 2014; 9(10):e106455. https://doi.org/10.1371/journal.pone.0106455 PMID: 25271417

40. Timp JF, Braekkan SK, Lijfering WM, van Hylckama Vlieg A, Hansen JB, Rosendaal FR, et al. Prediction of recurrent venous thrombosis in all patients with a first venous thrombotic event: The Leiden Thrombosis Recurrence Risk Prediction model (L-TRRiP). PLoS Med. 2019; 16(10):e1002883. https://doi.org/10.1371/journal.pmed.1002883 PMID: 31603898

41. Grant SW, Hickey GL, Grayson AD, Mitchell DC, McCollum CN. National risk prediction model for elective abdominal aortic aneurysm repair. Br J Surg. 2013; 100(5):645–53. https://doi.org/10.1002/bjs.9047 PMID: 23338659

42. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 2018; 10:100–7.

43. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018; 1:18. https://doi.org/10.1038/s41746-018-0029-1 PMID: 31304302

44. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). Heart. 2007; 93(2):172–6. https://doi.org/10.1136/hrt.2006.108167 PMID: 17090561

45. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007; 335(7611):136. https://doi.org/10.1136/bmj.39261.471806.55 PMID: 17615182

46. Batty GD, Gale CR, Kivimaki M, Deary IJ, Bell S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. BMJ. 2020; 368:m131. https://doi.org/10.1136/bmj.m131 PMID: 32051121

47. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. PeerJ. 2019; 7:e6257. https://doi.org/10.7717/peerj.6257 PMID: 30701130