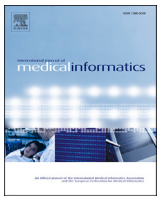
International Journal of Medical Informatics 179 (2023) 105209

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Cardiometabolic risk estimation using exposome data and machine learning

Angélica Atehortúa [a,*], Polyxeni Gkontra [a], Marina Camacho [a], Oliver Diaz [a], Maria Bulgheroni [b], Valentina Simonetti [b], Marc Chadeau-Hyam [c], Janine F. Felix [d,e], Sylvain Sebert [f], Karim Lekadir [a]

[a] BCN-AIM laboratory, Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
[b] R&D Ab.Acus s.r.l., Milano, Italy
[c] Department of Epidemiology and Biostatistics, MRC-HPA Centre for Environment and Health, School of Public Health, Imperial College London, London, United Kingdom
[d] The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[e] Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[f] Research Unit of Population Health, Faculty of Medicine, University of Oulu, Oulu, Finland

ARTICLE INFO

Keywords:
Exposure data
Cardiovascular disease
Type 2 diabetes
XGBoost
Explainability
Fairness

ABSTRACT

Background: The human exposome encompasses all exposures that individuals encounter throughout their lifetime. It is now widely acknowledged that health outcomes are influenced not only by genetic factors but also by the interactions between these factors and various exposures. Consequently, the exposome has emerged as a significant contributor to the overall risk of developing major diseases, such as cardiovascular disease (CVD) and diabetes. Therefore, personalized early risk assessment based on exposome attributes might be a promising tool for identifying high-risk individuals and improving disease prevention.

Objective: Develop and evaluate a novel and fair machine learning (ML) model for CVD and type 2 diabetes (T2D) risk prediction based on a set of readily available exposome factors. We evaluated our model using internal and external validation groups from a multi-center cohort. To be considered fair, the model was required to demonstrate consistent performance across different sub-groups of the cohort.

Methods: From the UK Biobank, we identified 5,348 and 1,534 participants who within 13 years from the baseline visit were diagnosed with CVD and T2D, respectively. An equal number of participants who did not develop these pathologies were randomly selected as the control group. 109 readily available exposure variables from six different categories (physical measures, environmental, lifestyle, mental health events, sociodemographics, and early-life factors) from the participant's baseline visit were considered. We adopted the XGBoost ensemble model to predict individuals at risk of developing the diseases. The model's performance was compared to that of an integrative ML model which is based on a set of biological, clinical, physical, and sociodemographic variables, and, additionally for CVD, to the Framingham risk score. Moreover, we assessed the proposed model for potential bias related to sex, ethnicity, and age. Lastly, we interpreted the model's results using SHAP, a state-of-the-art explainability method.

Results: The proposed ML model presents a comparable performance to the integrative ML model despite using solely exposome information, achieving a ROC-AUC of 0.78 ± 0.01 and 0.77 ± 0.01 for CVD and T2D, respectively. Additionally, for CVD risk prediction, the exposome-based model presents an improved performance over the traditional Framingham risk score. No bias in terms of key sensitive variables was identified.

Conclusions: We identified exposome factors that play an important role in identifying patients at risk of CVD and T2D, such as naps during the day, age completed full-time education, past tobacco smoking, frequency of tiredness/unenthusiasm, and current work status. Overall, this work demonstrates the potential of exposome-based machine learning as a fair CVD and T2D risk assessment tool.

1. Introduction

Risk assessment is essential in the prevention of high-burden diseases, such as cardiovascular disease (CVD) [1–3] and type 2 diabetes

* Corresponding author.
E-mail address: amatehortual@ub.edu (A. Atehortúa).

https://doi.org/10.1016/j.ijmedinf.2023.105209
Received 13 April 2023; Received in revised form 11 August 2023; Accepted 30 August 2023
Available online 12 September 2023

(T2D) [4,5]. It has the potential to promote the application of preventive measures, such as beneficial lifestyle change, and enhance adherence to medical advice [6,7]. The importance of early and personalized risk assessment has led to the development of a plethora of tools in the cardiovascular domain. These include the Framingham risk score [8], QRISK [9–11], the American Heart Association/Atherosclerotic Cardiovascular Disease (AHA/ASCVD) risk score calculator [12], AUDIRISK [13], to name a few. Despite their importance, these tools are based on simple algorithms assuming linearity, such as the Cox regression model [14], while they consider a limited number of traditional cardiovascular risk factors. Compared to these approaches, machine learning (ML) can handle vast amounts of numerical and categorical data without assumptions regarding the nature of the data such as normality or linearity. Therefore, it is emerging as an alternative for risk prediction that will allow the exploitation of additional data to achieve early and personalized identification of individuals at high risk of CVD and diabetes [15,16].

Exposure data, i.e. environmental, lifestyle, and behavioral factors, are of particular interest as potential predictors for such novel ML risk assessment models due to the information richness regarding the health of individuals that they encompass [17]. The contribution of environmental factors to disease risk is estimated to be 70 to 90% [18–20]. Exposures can be divided into external and internal. Typical external exposures include toxicants in the general environment and workplaces, diet, lifestyle, academic level, and socioeconomic status, while internal exposures are related to biological factors such as metabolic factors, gut microflora, and inflammation. All the exposures of an individual across the life course are called the exposome [21]. The exposome is complementary to the genome and can provide an improved understanding of the relation between risk factors and diseases leading to better prevention of chronic diseases [22]. Unlike clinical and radiological data [23], external exposome data are relatively easy to acquire using questionnaires or sensors located in smartphones, computers, or any electronic wearable device. The easiness of acquisition of exposome data allows for obtaining large volumes of individual-specific information that can be used with ML to obtain unprecedented insights into disease and enhance risk assessment.

Recently, several studies have used machine learning and statistical techniques to predict T2D and CVD based on different combinations of the aforementioned categories of predictors, as summarized in Table 1. More precisely, in the cardiovascular domain, Alaa et al. [24] predicted CVD risk by using ensemble ML models with data from nine categories: health and medical history, lifestyle and environment, blood assays, physical activity, family history, physical measures, psychosocial factors, dietary and nutritional information, and sociodemographics. Widen et al. [25] trained predictors for blood and urine markers (e.g. high-density lipoprotein, low-density lipoprotein, lipoprotein A, glycated hemoglobin, etc.) from single nucleotide polymorphisms (SNP) genotype to predict the CVD risk. Other works detected CVD from cardiac magnetic resonance imaging (CMR) phenotypes and genetic data by means of ML and Mendelian randomization [26–28].

In the study of T2D, associations between T2D and ready-to-eat food environments were identified by applying logistic regression on exposure data [29]. Lam et al. [30] developed diabetes prediction models using ML algorithms on data from wearable activity sensors, specifically wrist-worn triaxial accelerometers. Doleza et al. [31] devised a T2D risk prediction model using deep learning and features obtained by a smartphone, including demographic characteristics, anthropometric measures, lifestyle measures, medical history, and family history.

Despite the importance of these studies, most of the works included biological data, which might not always be available or easily accessible to the population, limiting their potential as self-assessment tools. Moreover, they neglected the use of exposome data or considered a very limited number of exposures, while traditional linear modeling techniques, such as logistic regression employed in these works, fail to optimally explore the richness of the exposome data to produce accurate risk estimations. To overcome these limitations, we propose a novel approach for identifying individuals at risk of CVD and T2D, respectively, based on easily accessible exposome factors and a state-of-art machine learning model, the XGBoost ensemble model [32]. The main contributions of this paper are summarized as follows:

- ·We present the first study to explore machine learning with a wide variety of exposome data, including physical measures, environmental, lifestyle, traumatic and psychosocial events, sociodemographics, and early-life factors, for CVD and T2D risk prediction using such a large and multi-center cohort as the UK Biobank (UKBB). The model is evaluated using internal and external validation based on data from independent assessment centers.
- ·Using the state-of-the-art explainability method, SHAP (SHapley Additive exPlanations), we have identified key exposome attributes in CVD and T2D risk prediction. These features might serve as potential risk factors for CVD and T2D and serve to build a rapid and accessible (self-)assessment risk prediction tool. Moreover, the knowledge gained regarding exposome CVD and T2D risk factors has the potential to drive the implementation of cost-effective preventive measures and policies to protect individuals' health from adverse environmental and lifestyle exposures [33].
- The proposed model is evaluated in terms of fairness regarding key sensitive variables (sex, ethnicity, age). We demonstrate that the model exhibits no bias across these variables bringing the model closer to real-world implementation.

2. Methods

An overview of the proposed approach is provided in Fig. 1. Each step of the pipeline is presented in detail in the following sections.

2.1. Data

Population Data from the UKBB application 65769 was used. The UKBB comprises data from 502,664 participants recruited from the UK National Health Service between 2006-2019 aged between 37 and 73 years. The participants have realized up to four assessment visits: a baseline visit (2006-2010), a first repeat visit (2012-2013), an imaging visit (2014+), and a first repeat imaging visit (2019+). During the first assessment visit, participants reported physical measurements, lifestyle, environmental, sociodemographic, traumatic, and physiological events, early-life factors as well as medical history. In subsequent visits, the information regarding medical history was updated. The participants consented to provide this information using a computer-based questionnaire under the ethical approval granted to Biobank from the Research Ethics Committee - REC reference 11/NW/0382 [34].

For the aim of this study, we used the exposome information gathered during the first assessment visit from participants without CVD or T2D to predict the development of the respective pathology as reported in subsequent visits. It should be noted that visits were carried out in 19 different clinical centers in the UK (Supplementary Table 1). Data from 3 different, independent centers than those used for training and internal validation, were used for external validation. The study cohort selection process is presented in Fig. 2. In brief, we enrolled in our study two groups of participants based on the ICD-10 diagnosis codes (see Outcomes definition) to identify those at risk of CVD and TD2, separately: i) a diseased group consisting of participants with the pathology of interest diagnosed within 13 years after the baseline visit (CVD or TD2), and ii) a control group consisting of an equal number of randomly selected participants that are healthy or suffering from other diseases than the diseases under study. This process resulted in a total of 5,348 with CVD and an equal number of participants without CVD being used for CVD prediction. Data from participants from 16 centers were used for training and internal validation (4,829 participants per

Table 1
Predictors used for risk assessment by different state-of-the-art methods. Categories included: Genetic data (GEN); medical imaging (IMG); biomarkers (BIOM) such as blood pressure, neuroticism score, and blood assays; environmental (ENV); lifestyle (LIFE); sociodemographic (SOCIOD); mental health (MH); physical measures (PHY) and early-life (EALI). The last column (UKBB) indicates whether the work used the UK biobank data.

Disease	Publication	Category of predictors									UKBB
		GEN	IMG	BIOM	ENV	LIFE	SOCIOD	MH	PHY	EALI	
CVD	Alaa et al. [24]	✗	✗	✓	✗	✓	✓	✗	✓	✗	✓
	Widen et al. [25]	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓
	Zheng et al. [26,27]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓
	Li et al. [28]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
T2D	Sarkar et al. [29]	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
	Lam et al. [30]	✗	✗	✗	✗	✓	✗	✗	✓	✗	✓
	Doleza et al. [31]	✗	✗	✓	✗	✓	✓	✗	✓	✗	✓
CVD, T2D	Proposed approach	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓

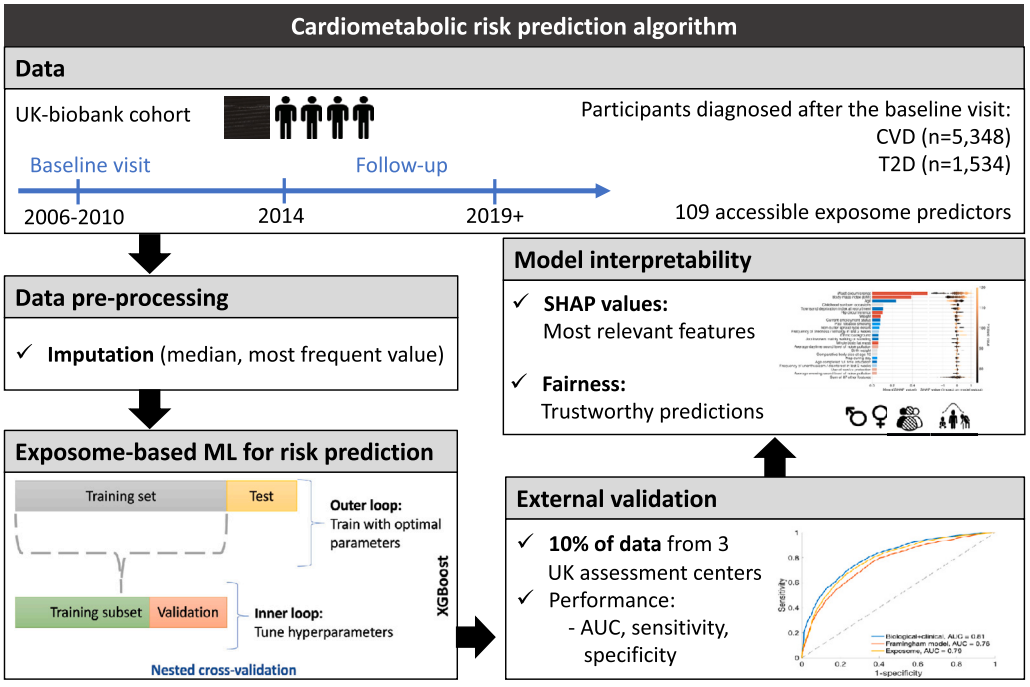


Fig. 1. Overview of the proposed approach for identifying individuals at risk of CVD and T2D using exposome data and machine learning (ML).

class, i.e. control group and individuals at risk of CVD), while data from the remaining 3 centers were used for external validation (519 participants per class). Similarly, 1,534 participants with TD2 and an equal number of participants without T2D were used for T2D prediction. From those participants, 1,390 per class were used for training and internal validation and 144 were kept for external validation.

Outcomes definition Participants with CVD were considered those with any of the following ICD-10 codes: coronary/ischaemic heart diseases (I20-I25), heart failure events (I50), vascular dementia (F01), and cerebrovascular diseases (I60-I69). For identifying participants with T2D, we included participants suffering from insulin-dependent (E10), non-insulin-dependent diabetes mellitus (E11), malnutrition-related diabetes mellitus (E12), and other specified diabetes mellitus (E13).

2.2. Data pre-processing

In total, we considered 109 exposome features from six different categories: early-life (14 features), environmental factors (9 features), lifestyle (46 features), sociodemographics (13 features), mental health (18 features), and physical measures (9 features). The complete list of features used in this work is provided in Supplementary Table 2.

Pre-processing was performed to curate the dataset for missing values. More precisely, we excluded from the study participants with 90% or above of missing data and those who answered “prefer not to answer” or “do not know” to any question, due to it can be considered as a missing value. This resulted in a sample of 1,534 individuals with T2D and 5,348 with CVD respectively, diagnosed after the baseline visit. An equal number of participants without the diseases under study was included in the control group. After removing participants, we imputed missing values by replacing them with the median and the most frequent value for numerical and categorical data, respectively. To choose the imputation approach, we performed a sensitivity analysis including a popular, more sophisticated imputation method, the Missforest algorithm [35]. The analysis consisted in comparing models with imputed data using MissForest or simple imputation with different percentages of missing values (15%, 20%, 25%, 30%). A paired t-test on the distributions of area under the curve (AUC) performances was performed in the nested cross-validation framework. No significant

Table 2
Participants' baseline characteristics used for disease prediction (internal|external validations). CVD and T2D stand for cardiovascular disease and diabetes, respectively. The number of individuals (n) and their mean and standard deviation (SD) are reported.

Characteristics	CVD ($n = 4,829\|519$)	Control without CVD ($n = 4,829\|519$)	T2D ($n = 1,390\|144$)	Control without T2D ($n = 1,390\|144$)
Age, years [mean(SD)]	74.9(6.6)|74.8(7.2)	69.9(8.0)|70.0(8.5)	73.4(7.3)|73.8(7.6)	70.2(8.2)|70.4(8.0)
Sex [n(%)]				
- Male	3,175(65.7)|333(64.2)	2,092(43.3)|227(43.8)	836(60.1)|87(60.4)	642(46.2)|69(47.9)
- Female	1,654(34.3)|186(35.8)	2,737(56.7)|292(56.2)	554(39.9)|57(39.6)	748(53.8)|75(52.1)
BMI, Kg/m^2 [mean(SD)]	28.9(5.0)|28.9(5.4)	27.2(4.7)|27.2(4.9)	31.5(5.6)|31.1(5.7)	27.1(4.6)|27.4(4.9)
Ethnicity [n(%)]				
- White	4,584(94.9)|479(92.3)	4,570(94.6)|480(92.4)	1,236(88.9)|134(93.1)	1,317(94.6)|140(97.2)
- Mixed	22(0.5)|2(0.4)	26(0.6)|5(1.0)	7(0.5)|2(1.4)	6(0.4)|1(0.7)
- Asian	137(2.8)|11(2.1)	102(2.1)|6(1.2)	74(5.3)|5(3.5)	27(1.9)|1(0.7)
- Black	51(1.1)|15(2.9)	73(1.5)|20(3.9)	47(3.5)|1(0.7)	23(1.6)|1(0.7)
- Chinese	5(0.1)|2(0.4)	14(0.3)|2(0.3)	6(0.4)|0(0.0)	3(0.5)|0(0.0)
- Other group	30(0.6)|10(1.9)	44(0.9)|6(1.2)	18(1.4)|2(1.4)	14(1.0)|1(0.7)
Age completed education, years [mean(SD)]	16.2(2.2)|16.2(2.5)	16.5(2.2)|16.4(2.0)	16.3(2.1)|16.0(1.8)	16.5(1.9)|16.6(3.0)
Current employment [n(%)]				
- Employed	1,792(37.1)|189(36.4)	2,922(60.5)|330(63.6)	591(42.5)|68(47.2)	838(60.3)|94(65.3)
- Retired	2,487(51.5)|241(46.4)	1,572(32.6)|132(25.4)	614(44.2)|60(41.7)	458(32.9)|42(29.2)
- Looking after home	65(1.3)|4(0.8)	137(2.8)|16(3.1)	36(2.6)|5(3.5)	29(2.1)|1(0.7)
- Unable to work	372(7.7)|64(12.3)	111(2.3)|23(4.4)	109(7.8)|10(6.9)	43(3.1)|6(4.2)
- Unemployed	91(1.9)|17(3.3)	64(1.3)|10(1.9)	36(2.6)|0(0.0)	17(1.2)|1(0.7)
- Doing unpaid work	14(0.3)|3(0.6)	13(0.3)|3(0.6)	3(0.2)|1(0.7)	4(0.3)|0(0.0)
- Full or part-time student	8(0.2)|1(0.2)	10(0.2)|5(1.0)	1(0.1)|0(0.0)	1(0.1)|0(0.0)

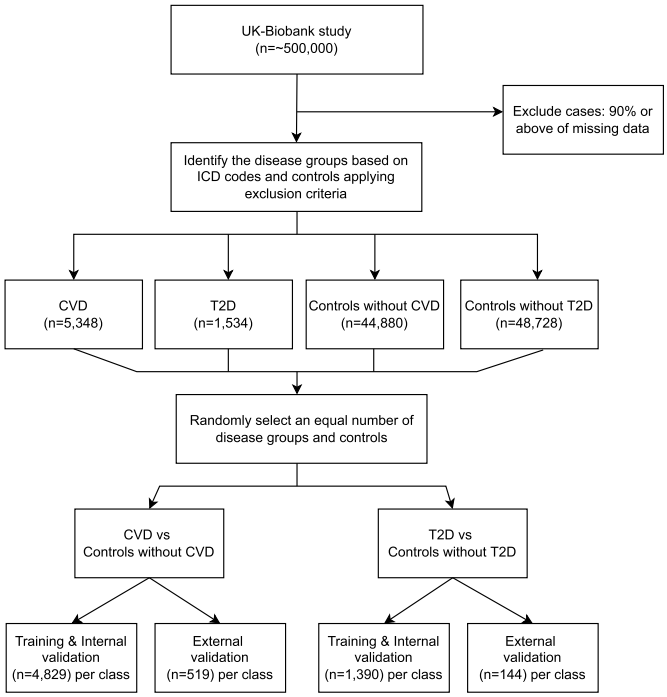


Fig. 2. Study cohorts selection process. Cardiovascular disease (CVD), Diabetes mellitus (T2D).

difference ($p - value > 0.05$) was observed when using missForest and simple imputation, so we adopted the second approach.

Fig. 2 presents the individuals' selection process, and Table 2 shows the population's characteristics.

2.3. Exposome-based machine learning for risk prediction

We used XGBoost [32] for identifying the participants at risk of developing CVD and T2D. XGBoost is a robust machine learning algorithm that achieves high predictive accuracy for high-dimensional problems with heterogeneous data (numerical and categorical) and missing values. It is an ensemble learning method that reaches a decision by combining the outputs from individual decision trees.

2.4. Internal and external validation

To evaluate the performance of our model we used internal and external validation groups. The external validation group comprised participants from 3 independent assessment centers selected randomly, corresponding to 10% of the study population. We studied if the selection of centers affected the performance of the models, and no significant difference was found by testing the model in different hold-out centers. The internal validation group comprised the participants from the 19 remaining centers.

Nested cross-validation with seven outer folds was used to evaluate the ML algorithm's performance in the internal validation group. On each of the outer folds, the training fold was divided into 5 inner cross-validation folds for hyper-parameter tuning. Grid search was used to identify the optimal set of parameters and select the best model for each outer fold. The selected model was then evaluated in the external validation group. XGBoost's learning rate, minimum child weight, gamma value, subsample, subsampling of columns per tree, and maximum depth parameters were selected from values: [0.05, 0.10, 0.15, 0.20, 0.25, 0.30], [1, 5, 10], [0.5, 1, 5], [0.6, 0.8, 1.0], [0.6, 0.8, 1.0], [3, 4, 5], respectively. The XGBoost model was implemented in Python 3.8 using the Scikit-Learn library 1.0 [36]. The source code of this work is available at https://github.com/amatehortual18/Cardiometabolic-risk-prediction-with-machine-learning. A comprehensive list of the fields of the UK Biobank used to develop the proposed machine learning model is provided within our GitHub repository. Interested researchers can apply for access to the fields by submitting a request to the UK Biobank.

To assess the performance of the proposed model and compare it with the reference models we used sensitivity, specificity, precision, and area under the receiver operating characteristic curve (AUC).

2.5. Model interpretability

One of the main contributions of this work is the identification of potentially modifiable exposome attributes that play a key role in CVD and T2D risk prediction. These attributes could be targeted for lifestyle and exposure interventions focused on the prevention of CVD and diabetes. Toward this aim, we determined the more important features by extracting the SHAP (SHapley Additive exPlanations) values [37,38]. This method is based on cooperative game theory to determine the importance of a single feature by computing the average contribution of that feature to the predictions across all possible feature combinations. More precisely, a value, called SHAP value, is assigned to each attribute included in the predictive model, based on the change in the prediction when a specific feature is included or excluded, taking into account the interactions and dependencies between features. Therefore, these SHAP values provide an explanation of the contribution of each feature to a particular prediction consistently and fairly, giving insights into how the ML model arrived at its decision. SHAP values provide individual-level interpretability as opposed to target-encoder techniques that are focused on encoding the target variable's statistics at a group level, providing insights into the average behavior of groups rather than individual predictions.

3. Results

3.1. Comparison to existing models

We compared our results to an ML model that uses biological, clinical, physical, and sociodemographic predictors based on the recent work by Alaa et al. [24]. In the remainder of the paper, we will refer to this model as the biological+clinical model. For the proposed risk models, we used 109 exposure factors, while excluding medical information, biomarkers such as blood assays, and other variables not easily accessible in daily life (i.e. diastolic and systolic blood pressure, impedance, arm fat mass, treatment/medication). Moreover, we included 54 exposome factors not considered in the existing biological+clinical model, such as environmental, early-life, and mental health factors. We focused on using only accessible exposome data to estimate the disease risk in a more personalized way. These factors are easily interpretable and some of them are modifiable.

Moreover, for CVD risk prediction, we compared the performance of our algorithm to that of a well-established model, the Framingham risk score [39]. The Framingham score is based on a set of conventional risk factors: age, sex, LDL cholesterol, HDL cholesterol, systolic blood pressure, diabetes, and smoking. Here, we computed the traditional Framingham score with the mentioned variables and furthermore, these factors were used to train an XGboost ML model to predict the CVD risk, using the same experimental setup described for our model.

3.2. Individuals at risk of CVD

At each fold, 3,863 subjects per class, i.e. control group and individuals at risk of CVD, were used to train the machine learning model. The remaining 966 and 519 subjects were used for internal and external validation, respectively. The optimal hyperparameters leading to the best model (biological+exposome) performance were a learning rate of 0.15, a minimum child weight of 5, a gamma value of 0.5, a subsampling rate of 0.8, a column subsampling rate of 0.8 per tree, and a maximum depth of 4.

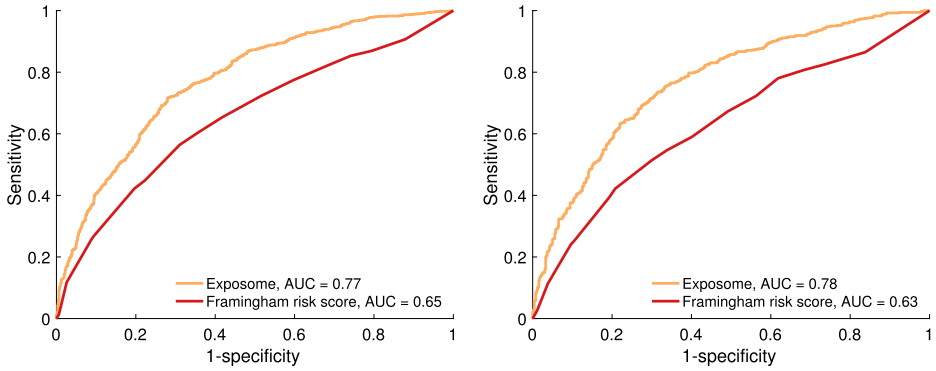
Table 3 allows for comparing in terms of sensitivity, specificity, precision, and AUC the performance of (i) the proposed ML exposome-based model, (ii) the biological+clinical model based on the work of Alaa et al. [24], (iii) the traditional Framingham risk score [39], (iv) an ML model based on XGBoost and the Framingham variables, and (v) an XGBoost model comprising all features used in previous models (exposome, clinical and biological information. Using the proposed exposome-based model, we achieved an AUC of 77|78% for identifying individuals at risk of CVD in the internal|external validation, corresponding to a statistically significant improvement ($p - value = 0.01$) over the traditional Framingham score that has very limited performance in this cohort (AUC of 66|64%). Fig. 3 presents a comparison between the AUC of the two approaches. The proposed model also outperforms the ML model based on the Framingham variables (XGBoost) by 2|8% with the difference being statistically significant ($p - value = 0.04$). Moreover, the proposed model, despite being solely based on exposome data, presented a comparable behavior to the Alaa et al. model with no statistically significant difference in the performance of the two models in terms of AUC ($p - value > 0.05$). Lastly, we used all features, i.e. exposome, clinical and biological information available, to assess the performance of a more complete model. This model outperformed all models with an AUC of 82% and 80% for internal and external validation, respectively. However, this model cannot be used for self-assessment, while prescribed medications, an actual indicator of the diseases, have the highest weight for the CVD prediction and, therefore, are the main contributors to the high model's performance.

In addition, we evaluated the CVD risk prediction at 5, 9, and 13 years. The results are provided in Table 4. The exposome-based model is able to predict the CVD risk in 5, 9, and 13 years with an AUC in the range of 75%-79% (see Fig. 4), achieving the highest performance for shorter-term predictions, i.e. 5 and 9 years. Interestingly, the exposome-based model outperforms the biological+clinical and Framingham model (XGBoost) in the external validation cohorts in terms of sensitivity, an important performance metric in cases where prediction of the individuals at high risk is of higher priority than specificity. Moreover, note how AUC obtained from the exposome-based model presents a lower variability at different prediction times (up to 5, 9, and 13 years), in comparison with the other models (biological+clinical) and Framingham model (XGBoost)). This fact implies a higher stability of the exposome-based model for the prediction of CVD risk.

The 22 most important variables involved in the CVD risk prediction using the exposome-based model are presented in Fig. 5. Please note that the number of factors to be displayed was selected after experimentation as a trade-off between clear visualization, feature importance, and the category of the exposome factors. However, the importance of all factors was calculated, and a complete list is provided in Supplementary Fig. 1. Sociodemographic factors, such as employment status, material deprivation as quantified by the Townsend index, qualifications, nap during the day, and age completed full-time education had a high impact on the CVD prediction. Moreover, lifestyle choices such as dietary habits, sleep duration, coffee type (decaffeinated, instant, ground), and tobacco were associated with CVD risk. This is in agreement with findings from previous studies [40,41]. Notably, factors related to mental health, such as frequency of tiredness and tenseness are among the factors that contribute the most to identifying individuals at risk of developing CVD. Please note that SHAP allows quantifying the overall effect of each feature by means of the mean absolute SHAP value (left panel, Fig. 5), but also the direction of the impact of the features on the model output (right panel of Fig. 5). For example, high frequency of tiredness values has a high positive contribution to the prediction, while lower values of this variable have a high negative contribution, indicating that frequency of tiredness is positively associated with the risk for CVD.

3.3. Individuals at risk of diabetes mellitus (T2D)

Similarly to CVD, we trained our model to identify individuals at risk of T2D. To this end, at each fold, we used 1,112 participants diagnosed with T2D and an equal number of control participants. 278 and 144 individuals per class were used for internal and external validations for each fold, respectively. The optimal parameters obtained for the higher results (biological+exposome model) were a learning rate of 0.10, a minimum child weight of 5, a gamma value of 0.5, a subsample rate of

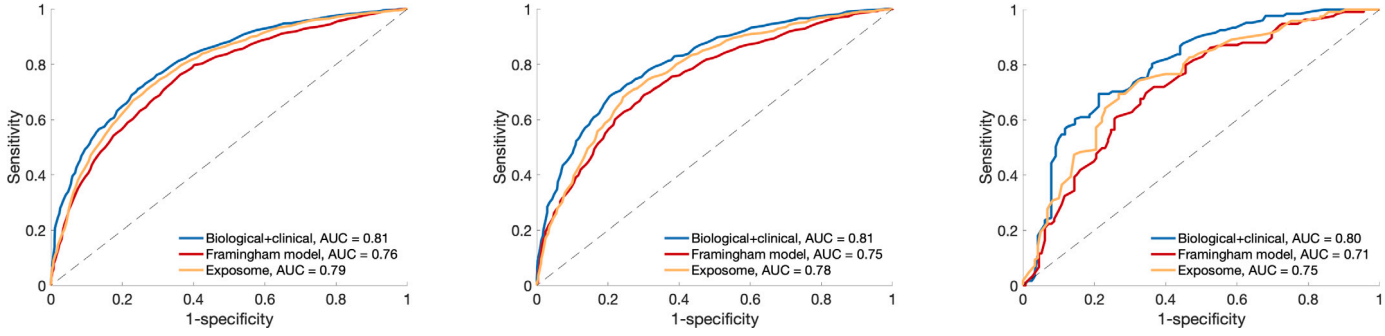


(a) Internal validation (b) External validation

Fig. 3. ROC AUC for predicting the CVD risk by means of the proposed exposome-based ML model (red) and the traditional Framingham risk score (orange). Results are presented for the (a) internal, and (b) external validation groups. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 3
CVD risk prediction using XGBoost. Mean and standard deviation results are shown from internal|external validation.

Model	Sensitivity	Specificity	Precision	AUC
Framingham score	$0.43 \pm 0.01 \vert 0.40 \pm 0.02$	$\mathbf{0.80 \pm 0.01 \vert 0.82 \pm 0.05}$	$0.69 \pm 0.01 \vert 0.69 \pm 0.03$	$0.66 \pm 0.01 \vert 0.64 \pm 0.02$
Framingham score (XGBoost)	$0.66 \pm 0.01 \vert 0.67 \pm 0.03$	$0.70 \pm 0.02 \vert 0.71 \pm 0.04$	$0.69 \pm 0.01 \vert 0.69 \pm 0.03$	$0.75 \pm 0.05 \vert 0.63 \pm 0.05$
Biological+clinical	$0.70 \pm 0.02 \vert 0.71 \pm 0.02$	$0.77 \pm 0.02 \vert 0.73 \pm 0.01$	$\mathbf{0.75 \pm 0.02 \vert 0.72 \pm 0.02}$	$0.81 \pm 0.02 \vert \mathbf{0.80 \pm 0.02}$
Exposome	$\mathbf{0.72 \pm 0.01 \vert 0.75 \pm 0.01}$	$0.68 \pm 0.02 \vert 0.66 \pm 0.01$	$0.70 \pm 0.01 \vert 0.69 \pm 0.01$	$0.77 \pm 0.01 \vert 0.78 \pm 0.01$
Exposome+Biological+clinical	$0.72 \pm 0.02 \vert 0.71 \pm 0.02$	$0.76 \pm 0.02 \vert 0.73 \pm 0.01$	$\mathbf{0.75 \pm 0.02} \vert 0.71 \pm 0.01$	$\mathbf{0.82 \pm 0.02 \vert 0.80 \pm 0.02}$



(a) 5 years (b) 9 years (c) 13 years

Fig. 4. ROC AUC for identifying people at risk of CVD within (a) 5, (b) 9, and (c) and 13 years using the biological+clinical model (blue), the Framingham model-XGBoost (red), and the proposed exposome-based model (yellow). Results are presented for the external validation group. The gray dashed line represents the ROC AUC curve for a random guess.

Table 4
Mean and standard deviation of AUC, precision, sensitivity, and specificity for CVD risk prediction over time in the internal | external validation groups.

Metric	Model	5 years ($n = 598 \vert 328$)	9 years ($n = 318 \vert 166$)	13 years ($n = 50 \vert 25$)
AUC	Framingham model (XGBoost)	$0.75 \pm 0.01 \vert 0.76 \pm 0.03$	$0.74 \pm 0.01 \vert 0.75 \pm 0.03$	$0.78 \pm 0.04 \vert 0.71 \pm 0.06$
	Biological+clinical	$0.82 \pm 0.01 \vert 0.81 \pm 0.02$	$0.80 \pm 0.02 \vert 0.81 \pm 0.02$	$0.81 \pm 0.04 \vert 0.80 \pm 0.02$
	Exposome	$0.77 \pm 0.01 \vert 0.79 \pm 0.01$	$0.77 \pm 0.01 \vert 0.78 \pm 0.02$	$0.78 \pm 0.02 \vert 0.75 \pm 0.05$
	Exposome+Biological+clinical	$0.83 \pm 0.01 \vert 0.81 \pm 0.03$	$0.81 \pm 0.01 \vert 0.77 \pm 0.03$	$0.82 \pm 0.03 \vert 0.80 \pm 0.03$
Precision	Framingham model (XGBoost)	$0.69 \pm 0.02 \vert 0.70 \pm 0.02$	$0.69 \pm 0.01 \vert 0.70 \pm 0.04$	$0.69 \pm 0.04 \vert 0.66 \pm 0.07$
	Biological+clinical	$0.76 \pm 0.02 \vert 0.75 \pm 0.02$	$0.75 \pm 0.03 \vert 0.77 \pm 0.01$	$0.80 \pm 0.08 \vert 0.71 \pm 0.04$
	Exposome	$0.69 \pm 0.01 \vert 0.71 \pm 0.02$	$0.71 \pm 0.02 \vert 0.73 \pm 0.03$	$0.69 \pm 0.05 \vert 0.69 \pm 0.05$
	Exposome+Biological+clinical	$0.75 \pm 0.01 \vert 0.72 \pm 0.04$	$0.75 \pm 0.02 \vert 0.70 \pm 0.04$	$0.77 \pm 0.06 \vert 0.78 \pm 0.03$
Sensitivity	Framingham model (XGBooost)	$0.67 \pm 0.02 \vert 0.67 \pm 0.02$	$0.64 \pm 0.03 \vert 0.66 \pm 0.03$	$0.72 \pm 0.06 \vert 0.64 \pm 0.14$
	Biological+clinical	$0.71 \pm 0.02 \vert 0.68 \pm 0.03$	$0.69 \pm 0.03 \vert 0.68 \pm 0.03$	$0.72 \pm 0.06 \vert 0.70 \pm 0.08$
	Exposome	$0.73 \pm 0.01 \vert 0.72 \pm 0.03$	$0.71 \pm 0.02 \vert 0.71 \pm 0.02$	$0.75 \pm 0.03 \vert 0.74 \pm 0.03$
	Exposome+Biological+clinical	$0.73 \pm 0.02 \vert 0.75 \pm 0.01$	$0.70 \pm 0.03 \vert 0.69 \pm 0.04$	$0.74 \pm 0.06 \vert 0.65 \pm 0.03$
Specificity	Framingham model (XGBoost)	$0.69 \pm 0.02 \vert 0.71 \pm 0.03$	$0.71 \pm 0.02 \vert 0.71 \pm 0.07$	$0.68 \pm 0.05 \vert 0.68 \pm 0.05$
	Biological+clinical	$0.77 \pm 0.02 \vert 0.77 \pm 0.04$	$0.77 \pm 0.03 \vert 0.79 \pm 0.01$	$0.81 \pm 0.08 \vert 0.72 \pm 0.04$
	Exposome	$0.68 \pm 0.02 \vert 0.71 \pm 0.01$	$0.70 \pm 0.02 \vert 0.73 \pm 0.05$	$0.67 \pm 0.06 \vert 0.68 \pm 0.05$
	Exposome+Biological+clinical	$0.76 \pm 0.01 \vert 0.72 \pm 0.04$	$0.77 \pm 0.02 \vert 0.70 \pm 0.03$	$0.79 \pm 0.06 \vert 0.78 \pm 0.03$

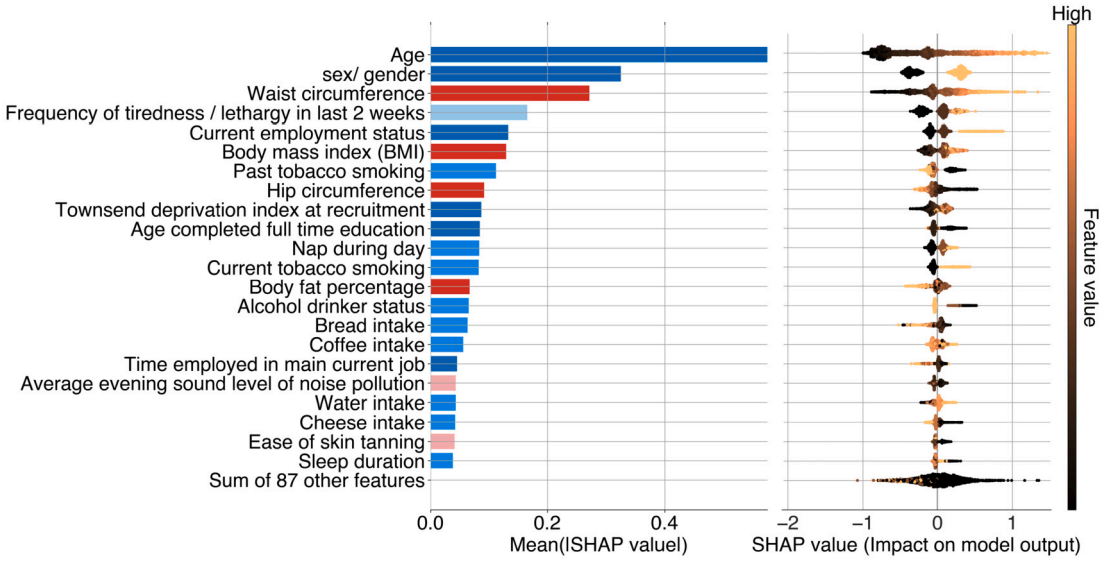


Fig. 5. Exposome-based model interpretability. The left panel provides the 22 most important factors involved in CVD risk prediction as ranked by the mean absolute SHAP value. The right panel shows the impact of each feature on the model output. Please note that each point represents a participant. Higher values of the features are indicated by orange, while lower with black. Exposome categories: Physical measures; Sociodemographics; Lifestyle; Mental health; Environmental.

Table 5
T2D risk prediction using XGBoost. Mean and standard deviation results are shown from internal|external validation.

Model	Sensitivity	Specificity	Precision	AUC
Biological+clinical	$0.78 \pm 0.02 \vert 0.71 \pm 0.02$	$0.73 \pm 0.04 \vert 0.77 \pm 0.01$	$0.74 \pm 0.03 \vert 0.75 \pm 0.02$	$0.82 \pm 0.02 \vert 0.81 \pm 0.01$
Exposome	$0.74 \pm 0.02 \vert 0.70 \pm 0.03$	$0.70 \pm 0.04 \vert 0.70 \pm 0.02$	$0.70 \pm 0.02 \vert 0.70 \pm 0.02$	$0.80 \pm 0.02 \vert 0.77 \pm 0.01$
Expsome+Biological+clinical	$0.78 \pm 0.01 \vert 0.72 \pm 0.02$	$0.73 \pm 0.02 \vert 0.75 \pm 0.02$	$0.74 \pm 0.02 \vert 0.74 \pm 0.02$	$0.83 \pm 0.02 \vert 0.81 \pm 0.01$

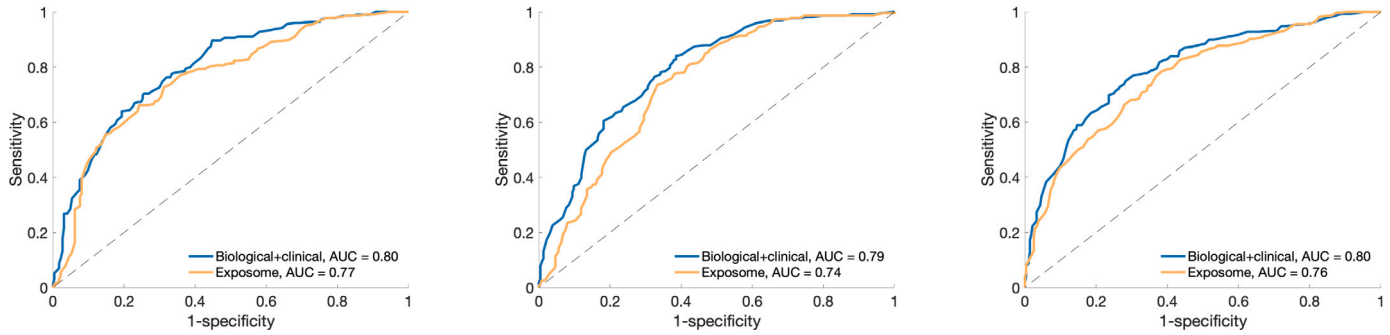
0.8, a column subsampling rate of 0.8 per tree, and a maximum depth of 5. Table 5 shows the predictive performance of the exposome-based ML model and the biological+clinical model based on Alaa et al. work trained to identify individuals at risk of diabetes. We obtain an AUC performance of around 80|77% during internal|external validation, which is 2|4% lower than the performance of the biological+clinical model ($p - value = 0.12$), whose variables (e.g. number of treatments/medications taken) require access to clinical services. Furthermore, the model performance was evaluated at different time points as presented in Table 6. We computed precision, sensitivity, specificity, and AUC based on the disease class, in which the diagnosis date is known. The exposome-based model is able to identify individuals at risk of T2D at 5, 9, and 13 years with an AUC in the range of 74%-80% (see Fig. 6). Despite the biological+clinical model achieving better performance in terms of AUC than the exposome-based model for different time points, the addition of exposome features to biological factors from blood assays provides a more reliable and precise prediction in 13 years in terms of AUC, which could be interesting for long-term detection of T2D. The exposome+biological+clinical model presented a slightly improved AUC by 1% in internal validation but did not outperform the biological+clinical model in external validation ($p - value = 0.26$).

Fig. 7 provides the 22 more important variables involved in the T2D risk prediction by means of the exposome-based model. Physical measures, such as waist/hip circumference, weight, and BMI contributed the most in identifying individuals at risk of T2D, with higher waist circumferences and BMI values being associated with a higher probability of developing T2D (Fig. 7, right panel). Moreover, sociodemographic factors, such as Townsend deprivation index, and those related to the individual's occupation (i.e. unemployment status from current employment and job involves mainly walking or standing) were associated with increased diabetes risk. These findings are aligned with results from previous research [42,43]. Using the exposome-based model, we were able to identify different categories of risk factors associated with T2D. These include diet-related habits, factors related to socioeconomic status (employment status, deprivation index), but also early-life features such as childhood sunburn occasions, birth weight, and body size at age ten. Furthermore, waist circumference, BMI, frequency of tiredness, and naps during the day were found associated with T2D. Many of the identified factors are closely linked to known T2D risk factors, such as socio-economic factors and early growth, or may represent known factors, e.g. sunburn may reflect time spent outside or be related to physical activity. Interestingly, we also identified environmental factors like average daytime sound level of noise pollution and sun exposure also related to the T2D risk. These modifiable factors can be quantitatively measured to predict the personalized risk of T2D at the time by the exposome-based model. On the contrary, for the biological+clinical model, as in the case of CVD, the clinical variables related to the number of treatments and biological data, such as white blood cell count, were the most important to identify individuals at risk of T2D, prohibiting the applicability of such of machine learning for fast assessment without the need of blood assays or as a self-assessment tool.

3.4. Exposome-based model with a reduced number of features

We also evaluated the performance of the exposome-based models for CVD and T2D, respectively using an increasing number of the most important features as provided by the Gini importance score of the model (Fig. 8). The results demonstrated that by using the 40 or 20 most important features our model already achieves similar performance (AUC, precision, and sensitivity) to using the entire set of exposome features for CVD and T2D prediction, respectively.

Algorithmic fairness Last but not least, we evaluated the fairness of the proposed models for CVD and T2D risk prediction by computing the statistical parity difference [44] and the disparate impact ratio [45]. The results are presented in Fig. 9 and 10. No bias was identified in



(a) 5 years (b) 9 years (c) 13 years

Fig. 6. ROC AUC for identifying people at risk of T2D within 5 (a), 9 (b), and 13 (c) years using the biological+clinical model (blue) and the proposed exposome-based model (yellow). Results are presented for the external validation cohort. The gray dashed line represents the ROC AUC curve for a random guess.

Table 6
Mean and standard deviation of precision, sensitivity, and specificity for T2D risk prediction over time for all clinical centers in the internal | external validations. The number of individuals, *n*, is shown per class.

Metric	Model	5 years ($n = 70\|47$)	9 years ($n = 100\|43$)	13 years ($n = 108\|54$)
AUC	Biological+clinical	$0.84 \pm 0.02\|0.80 \pm 0.04$	$0.83 \pm 0.03\|0.79 \pm 0.03$	$0.82 \pm 0.03\|0.80 \pm 0.03$
	Exposome	$0.80 \pm 0.02\|0.77 \pm 0.05$	$0.80 \pm 0.04\|0.74 \pm 0.04$	$0.79 \pm 0.02\|0.76 \pm 0.06$
	Exposome+Biological+clinical	$0.81 \pm 0.02\|0.79 \pm 0.01$	$0.84 \pm 0.02\|0.79 \pm 0.02$	$0.85 \pm 0.03\|0.85 \pm 0.01$
Precision	Biological+clinical	$0.77 \pm 0.03\|0.72 \pm 0.04$	$0.74 \pm 0.04\|0.70 \pm 0.03$	$0.75 \pm 0.03\|0.73 \pm 0.06$
	Exposome	$0.72 \pm 0.01\|0.69 \pm 0.02$	$0.71 \pm 0.05\|0.68 \pm 0.04$	$0.71 \pm 0.02\|0.68 \pm 0.07$
	Exposome+Biological+clinical	$0.73 \pm 0.02\|0.73 \pm 0.03$	$0.77 \pm 0.02\|0.70 \pm 0.02$	$0.75 \pm 0.01\|0.79 \pm 0.02$
Sensitivity	Biological+clinical	$0.79 \pm 0.01\|0.71 \pm 0.05$	$0.77 \pm 0.04\|0.75 \pm 0.05$	$0.77 \pm 0.06\|0.75 \pm 0.06$
	Exposome	$0.76 \pm 0.04\|0.74 \pm 0.07$	$0.73 \pm 0.05\|0.74 \pm 0.06$	$0.75 \pm 0.02\|0.73 \pm 0.03$
	Exposome+Biological+clinical	$0.75 \pm 0.01\|0.66 \pm 0.03$	$0.80 \pm 0.01\|0.77 \pm 0.03$	$0.79 \pm 0.01\|0.73 \pm 0.02$
Specificity	Biological+clinical	$0.77 \pm 0.04\|0.72 \pm 0.06$	$0.73 \pm 0.04\|0.69 \pm 0.03$	$0.74 \pm 0.02\|0.73 \pm 0.06$
	Exposome	$0.70 \pm 0.02\|0.67 \pm 0.05$	$0.70 \pm 0.07\|0.65 \pm 0.04$	$0.69 \pm 0.03\|0.65 \pm 0.10$
	Exposome+Biological+clinical	$0.72 \pm 0.01\|0.76 \pm 0.03$	$0.76 \pm 0.02\|0.67 \pm 0.03$	$0.73 \pm 0.01\|0.81 \pm 0.02$

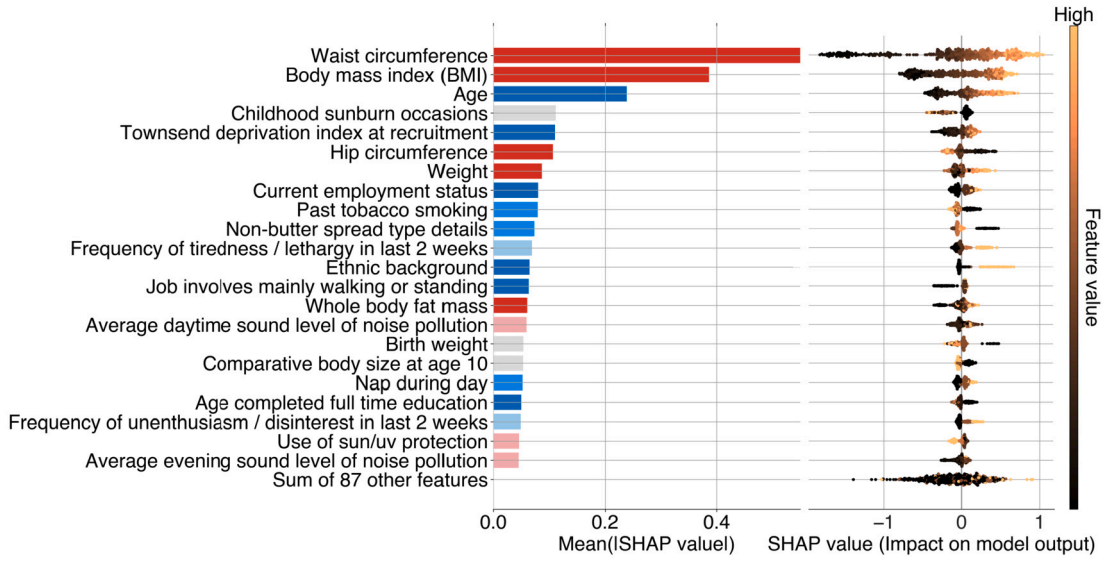
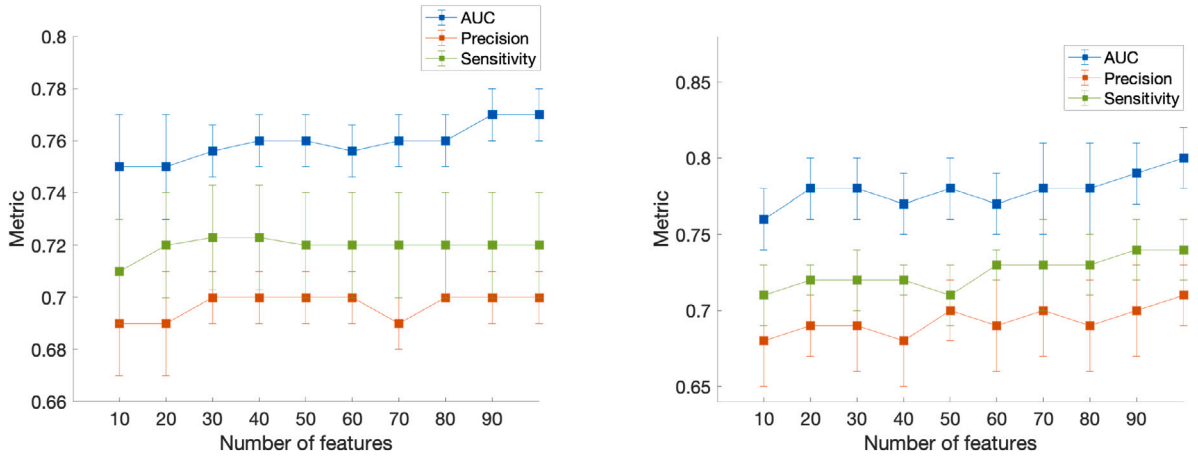
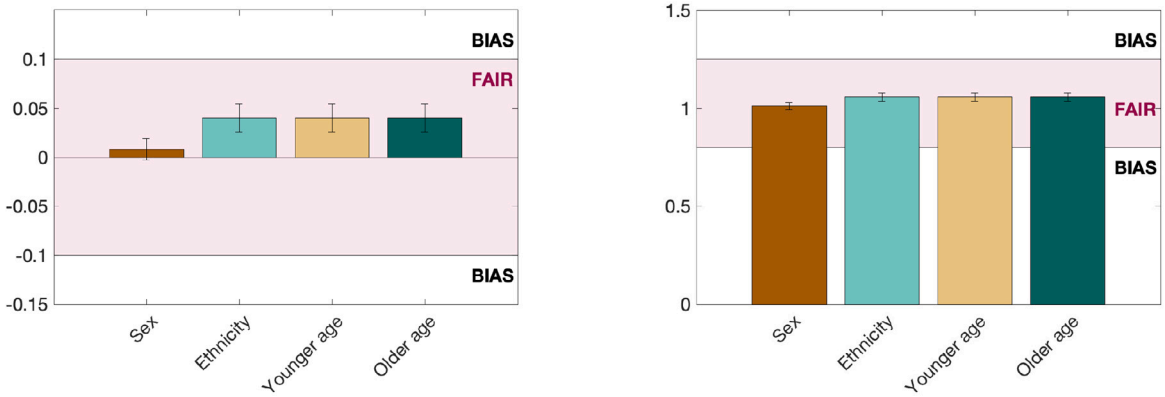


Fig. 7. Exposome-based model interpretability. The left panel shows the 22 most important factors involved in T2D risk prediction, which are expressed by the mean absolute SHAP. The right panel provides the impact of each feature on the model output. Please note that each point corresponds to a participant in the training set. Higher values of the features are indicated by orange, while lower with black. Exposome categories: Physical measures; Sociodemographics; Early-life; Lifestyle; Mental health; Environmental.



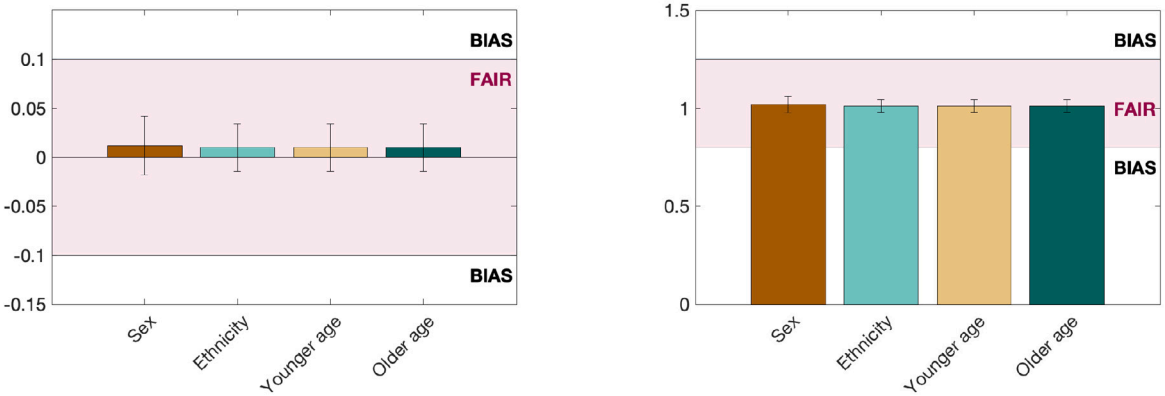
(a) CVD (b) T2D

Fig. 8. Exposome-based model performance considering an increasing number of features for (a) CVD, and (b) T2D risk prediction.



(a) Statistical parity difference (b) Disparate impact ratio

Fig. 9. Fairness performance of the exposome-based model for CVD prediction. Models are considered fair when having a statistical parity difference within the -0,1 and 01. Similarly, the fair models present disparate ratios between 0,75 and 1,25. Please note younger ages from 20 to 50 years.



(a) Statistical parity difference (b) Disparate impact ratio

Fig. 10. Fairness performance of the exposome-based model for T2D prediction.

terms of any of the considered sensitive variables (ethnicity, gender, or age). Please note that in order to assess age-related bias, two age groups were evaluated, a group of individuals of younger age (from 20 to 50 years) and a corresponding group of individuals of older age.

4. Discussion

Using a large longitudinal population cohort, the UKBB, we developed and validated two novel exposome-based ML risk prediction mod-

els to identify individuals at risk of two major diseases, CVD and T2D respectively. The proposed strategy combined exposome features from several different categories without including any clinical information that might be expensive, tedious, and time-consuming to acquire. The easily accessible exposome features highlight the potential of our model to be used as a tool for rapid assessment, including self-assessment.

A limited amount of research has employed exposome factors for CVD and T2D risk prediction (Table 1). Among those achieving the highest accuracy is the work of Alaa et al. [24], which included biological and clinical markers, as well as a few exposure predictors but without considering early-life factors for instance. By means of the proposed approach, we achieved comparable performance to this integrative model despite being solely based on readily available exposome factors, including early-life factors (AUC of 0.78 ± 0.01 and 0.77 ± 0.01 for CVD and diabetes, respectively, in external validation). For CVD risk prediction, the proposed exposome-based model also outperformed the widely-used Framingham risk score and an ML model that is based on the Framingham risk predictors. Furthermore, our findings demonstrated that for both CVD and T2D risk assessment, the most influential features are a blend of (i) sociodemographic features, including features related to work/education status, such as current employment status, age completed full-time education, (ii) physical measures, such as waist circumference, (iii) mental factors, such as frequency of tiredness and tenseness, and (iv) lifestyle factors, e.g. current tobacco smoking, nap during the day, alcohol drinker status and dietary habits. These findings are aligned with risk factors clinically reported in the literature [46,41]. Furthermore, we demonstrated that the exposome-based model is stable for a few number of features, as illustrated in Fig. 8. Therefore, a simpler version of the proposed CVD and T2D models using for example 40 or 20 features can be established to facilitate users by reducing the amount of information that they need to provide, without comprising significantly the performance.

In our study, we selected XGBoost as the preferred classification algorithm based on thorough experimentation with various state-of-the-art classification algorithms, including SVM, random forest, and AdaBoost. The results indicated that XGBoost outperformed the other algorithms. XGBoost has been widely recognized for its efficiency and effectiveness in diverse scenarios, outperforming even deep learning models for tabular data [47,48]. One of the key advantages of XGBoost is its ability to incorporate regularization techniques such as L1 and L2, preventing overfitting and improving generalization. Given that our exposome data encompasses heterogeneous information, we believe that the regularization capabilities of XGBoost played a crucial role in achieving the best performance in our study. Furthermore, XGBoost provided built-in feature importance estimation, identifying the most influential features in the complex exposome dataset, being aligned with the medical literature.

In addition, to minimize overfitting, in this work, we applied nested cross-validation for combined hyperparameter tuning and model selection based on the internal validation set. This procedure treats hyperparameter tuning as part of the model selection and evaluates it using an outer k-fold cross-validation. Therefore, it allows us to better estimate the model's generalization capability and identify whether overfitting occurs [49]. The results show stability without overfitting. Using the resulting optimal values for the depth of the trees (max depth) and learning rate in the XGBoost models as regularization terms simplify the ML model and reduce overfitting. Lastly, we evaluated the developed models in an external validation cohort. The results in the external validation were close to those obtained for the internal validation set demonstrating that the model generalizes well to unseen data, even from different centers.

Despite the importance of the findings, there exist some limitations in the present work. First, a high number of missing values existed in the UKBB cohort for some of the exposome variables. This is a common issue for longitudinal studies. To overcome this limitation, we used a model that handles well missing values, and, additionally, we adopted a simple, yet effective, imputation approach for categorical and numerical data [50]. Moreover, the present study was based on a predominantly white population from the UK. Future research using cohorts from different countries and ethnicities is needed to evaluate the generalizability and transferability of the proposed risk prediction model.

5. Conclusions

The exposome can modulate genetic effects, representing about 70 to 90% of the risk for major diseases [51,52]. Therefore, new strategies to predict the risk of major pathologies based on the human exposome represent an opportunity for improving early prevention and promoting beneficial lifestyle changes [33]. In this context, we performed the first study using a wide range of exposome data, including sociodemographic, lifestyle, environmental, occupational, psycho-social, mental, and early-life factors, to identify individuals at risk of two high-burden diseases; CVD and T2D. By leveraging machine learning and a large population cohort, we were able to exploit the wealth of information provided by the human exposome and demonstrate that an exposome-based machine learning model is a potentially powerful tool for accessible risk prediction in future healthcare in a personalized way.

Summary table

What was already known on the topic

- A large number of information is necessary to develop models for a personalized identification of the individuals at risk of diseases, however, the clinical information is usually limited to a few samples.
- Traditional clinical risk scores are based on general populations and linear algorithms without taking into account the potential heterogeneous information from an individual.

What this study added to our knowledge

- Our study resulted in an exposome-based machine learning model that predicts the risk of disease by using a large dataset and outperforms a well-established tool, the Framingham risk, and performs comparably to a more integrative model requiring clinical information
- As well, our machine learning model is interpretable and allows identifying key factors involved in the development of cardiovascular disease and type 2 diabetes
- Exposome data is potential information with a large number of samples that should support a personalized estimation of the disease risk
- The exposome-based models could be analyzed for early preventive measures such as beneficial lifestyle changes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement No 874739 (LongITools project). PG and KL have additionally received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825903 (euCanSHare project).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105209.

References

[1] X. Rossello, J.A. Dorresteijn, A. Janssen, E. Lambrinou, M. Scherrenberg, E. Bonnefoy-Cudraz, M. Cobain, M.F. Piepoli, F.L. Visseren, P. Dendale, T. paper is a co-publication betw, Risk prediction tools in cardiovascular disease prevention: A report from the ESC prevention of CVD programme led by the european association of preventive cardiology (EAPC) in collaboration with the acute cardiovascular care association (ACCA) and the association of cardiovascular nursing and allied professions (ACNAP), Eur. J. Prev. Cardiol. 26 (14) (2019) 1534–1544, https://doi.org/10.1177/2047487319846715.

[2] D.M. Lloyd-Jones, L.T. Braun, C.E. Ndumele, S.C. Smith Jr, L.S. Sperling, S.S. Virani, R.S. Blumenthal, Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the american heart association and american college of cardiology, Circulation 139 (25) (2019) e1162–e1177.

[3] M.S. Maron, E.J. Rowin, B.S. Wessler, P.J. Mooney, A. Fatima, P. Patel, B.C. Koethe, M. Romashko, M.S. Link, B.J. Maron, Enhanced american college of cardiology/american heart association strategy for prevention of sudden cardiac death in high-risk patients with hypertrophic cardiomyopathy, JAMA Cardiol. 4 (7) (2019) 644–657.

[4] B. Buijsse, R.K. Simmons, S.J. Griffin, M.B. Schulze, Risk assessment tools for identifying individuals at risk of developing type 2 diabetes, Epidemiol. Rev. 33 (1) (2011) 46–62, https://doi.org/10.1093/epirev/mxq019.

[5] H. Chatterton, T. Younger, A. Fischer, K. Khunti, Risk identification and interventions to prevent type 2 diabetes in adults at high risk: summary of NICE guidance, BMJ, Br. Med. J. 345 (jul12 3) (2012) e4624, https://doi.org/10.1136/bmj.e4624.

[6] R.F. Catalano, A.A. Fagan, L.E. Gavin, M.T. Greenberg, C.E. Irwin Jr, D.A. Ross, D.T. Shek, Worldwide application of prevention science in adolescent health, Lancet 379 (9826) (2012) 1653–1664.

[7] T.A. Pearson, G.A. Mensah, R.W. Alexander, J.L. Anderson, R.O. Cannon III, M. Criqui, Y.Y. Fadl, S.P. Fortmann, Y. Hong, G.L. Myers, et al., Markers of inflammation and cardiovascular disease: application to clinical and public health practice: a statement for healthcare professionals from the centers for disease control and prevention and the American heart association, Circulation 107 (3) (2003) 499–511.

[8] R.B. D'Agostino, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W.B. Kannel, General cardiovascular risk profile for use in primary care, Circulation 117 (6) (2008) 743–753, https://doi.org/10.1161/circulationaha.107.699579.

[9] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, P. Brindle, Derivation and validation of QRISK, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study, BMJ, Br. Med. J. 335 (7611) (2007) 136, https://doi.org/10.1136/bmj.39261.471806.55.

[10] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, R. Minhas, A. Sheikh, P. Brindle, Predicting cardiovascular risk in england and wales: prospective derivation and validation of QRISK2, BMJ, Br. Med. J. 336 (7659) (2008) 1475–1482, https://doi.org/10.1136/bmj.39609.449676.25.

[11] J. Hippisley-Cox, C. Coupland, P. Brindle, Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study, BMJ, Br. Med. J. (2017) j2099, https://doi.org/10.1136/bmj.j2099.

[12] D.K. Arnett, R.S. Blumenthal, M.A. Albert, A.B. Buroker, Z.D. Goldberger, E.J. Hahn, C.D. Himmelfarb, A. Khera, D. Lloyd-Jones, J.W. McEvoy, E.D. Michos, M.D. Miedema, D. Muñoz, S.C. Smith, S.S. Virani, K.A. Williams, J. Yeboah, B. Ziaeian, ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines, Circulation 140 (11) (2019), https://doi.org/10.1161/cir.0000000000000678.

[13] L. Chen, D.J. Magliano, B. Balkau, S. Colagiuri, P.Z. Zimmet, A.M. Tonkin, P. Mitchell, P.J. Phillips, J.E. Shaw, AUSDRISK: an australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures, Med. J. Aust. 192 (4) (2010) 197–202, https://doi.org/10.5694/j.1326-5377.2010.tb03507.x.

[14] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc., Ser. B, Methodol. 34 (2) (1972) 187–202, https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

[15] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, Disease prediction by machine learning over big data from healthcare communities, IEEE Access 5 (2017) 8869–8879.

[16] A. Poveda, H. Pomares-Millan, Y. Chen, A. Kurbasic, C.J. Patel, F. Renström, G. Hallmans, I. Johansson, P.W. Franks, Exposome-wide ranking of modifiable risk factors for cardiometabolic disease traits, Sci. Rep. 12 (1) (2022) 1–10.

[17] M. Vrijheid, The exposome: a new paradigm to study the impact of environment on health, Thorax 69 (9) (2014) 876–878.

[18] S.M. Rappaport, M.T. Smith, Environment and disease risks, Science 330 (6003) (2010) 460–461.

[19] R.V. Saveanu, C.B. Nemeroff, Etiology of depression: genetic and environmental factors, Psychiatr. Clin. 35 (1) (2012) 51–71.

[20] L. Maitre, M. Bustamante, C. Hernández-Ferrer, D. Thiel, C.-H.E. Lau, A.P. Siskos, M. Vives-Usano, C. Ruiz-Arenas, D. Pelegrí-Sisó, O. Robinson, et al., Multi-omics signatures of the human early life exposome, Nat. Commun. 13 (1) (2022) 7024.

[21] C.P. Wild, Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology, Cancer Epidemiol. Biomark. Prev. 14 (8) (2005) 1847–1850.

[22] D.G. DeBord, T. Carreón, T.J. Lentz, P.J. Middendorf, M.D. Hoover, P.A. Schulte, Use of the "exposome" in the practice of epidemiology: a primer on-omic technologies, Am. J. Epidemiol. 184 (4) (2016) 302–314.

[23] D.J. Park, M.W. Park, H. Lee, Y.-J. Kim, Y. Kim, Y.H. Park, Development of machine learning model for diagnostic disease prediction based on laboratory tests, Sci. Rep. 11 (1) (2021) 1–11.

[24] A.M. Alaa, T. Bolton, E. Di Angelantonio, J.H. Rudd, M. van der Schaar, Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 uk biobank participants, PLoS ONE 14 (5) (2019) e0213653.

[25] E. Widen, T.G. Raben, L. Lello, S.D. Hsu, Machine learning prediction of biomarkers from snps and of disease risk from biomarkers in the uk biobank, medRxiv.

[26] Q. Zheng, H. Delingette, K. Fung, S.E. Petersen, N. Ayache, Unsupervised shape and motion analysis of 3822 cardiac 4d mris of uk biobank, preprint, arXiv:1902.05811.

[27] Q. Zheng, H. Delingette, K. Fung, S.E. Petersen, N. Ayache, Pathological cluster identification by unsupervised analysis in 3,822 uk biobank cardiac mris, Front. Cardiovasc. Med. 7 (2020) 164.

[28] X. Li, X. Meng, Y. He, A. Spiliopoulou, M. Timofeeva, W.-Q. Wei, A. Gifford, T. Yang, T. Varley, I. Tzoulaki, et al., Genetically determined serum urate levels and cardiovascular and other diseases in uk biobank cohort: a phenome-wide mendelian randomization study, PLoS Med. 16 (10) (2019) e1002937.

[29] C. Sarkar, C. Webster, J. Gallacher, Are exposures to ready-to-eat food environments associated with type 2 diabetes? a cross-sectional study of 347 551 uk biobank adult participants, Lancet Planet. Health 2 (10) (2018) e438–e450.

[30] B. Lam, M. Catt, S. Cassidy, J. Bacardit, P. Darke, S. Butterfield, O. Alshabrawy, M. Trenell, P. Missier, Using wearable activity trackers to predict type 2 diabetes: machine learning–based cross-sectional study of the uk biobank accelerometer cohort, JMIR Diabet. 6 (1) (2021) e23364.

[31] N. Dolezalova, M. Cairo, A. Despotovic, A.T. Booth, A.B. Reed, D. Morelli, D. Plans, Development of a dynamic type 2 diabetes risk prediction tool: a uk biobank study, preprint, arXiv:2104.10108.

[32] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al., Xgboost: extreme gradient boosting, R package version 0.4-2 1 (4) (2015) 1–4.

[33] A.A. Baccarelli, The Human Exposome: A New "omic" Ready for Prime Time, 2019.

[34] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al., Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, PLoS Med. 12 (3) (2015) e1001779.

[35] D.J. Stekhoven, P. Bühlmann, Missforest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (1) (2012) 112–118.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[37] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[38] M.T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[39] D.M. Lloyd-Jones, P.W. Wilson, M.G. Larson, A. Beiser, E.P. Leip, R.B. D'Agostino, D. Levy, Framingham risk score and prediction of lifetime risk for coronary heart disease, Am. J. Cardiol. 94 (1) (2004) 20–24.

[40] T.K. Tegegne, S.M.S. Islam, R. Maddison, Effects of lifestyle risk behaviour clustering on cardiovascular disease among uk adults: latent class analysis with distal outcomes, Sci. Rep. 12 (1) (2022) 1–8.

[41] C. Méjean, M. Droomers, Y.T. Van Der Schouw, I. Sluijs, S. Czernichow, D.E. Grobbee, H.B. Bueno-de Mesquita, J.W. Beulens, The contribution of diet and lifestyle to socioeconomic inequalities in cardiovascular morbidity and mortality, Int. J. Cardiol. 168 (6) (2013) 5190–5195.

[42] S. Feller, H. Boeing, T. Pischon, Body mass index, waist circumference, and the risk of type 2 diabetes mellitus: implications for routine clinical practice, Dtsch. Arztebl. Int. 107 (26) (2010) 470.

[43] S. Carlsson, T. Andersson, M. Talbäck, M. Feychting, Incidence and prevalence of type 2 diabetes by occupation: results from all swedish employees, Diabetologia 63 (1) (2020) 95–103.

[44] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International Conference on Machine Learning, PMLR, 2013, pp. 325–333.

[45] S. Barocas, A.D. Selbst, Big data's disparate impact, Calif. Law Rev. (2016) 671–732.

[46] A.K. Chomistek, J.E. Manson, M.L. Stefanick, B. Lu, M. Sands-Lincoln, S.B. Going, L. Garcia, M.A. Allison, S.T. Sims, M.J. LaMonte, et al., Relationship of sedentary behavior and physical activity to incident cardiovascular disease: results from the women's health initiative, J. Am. Coll. Cardiol. 61 (23) (2013) 2346–2354.

[47] R. Shwartz-Ziv, A. Armon, Tabular data: deep learning is not all you need, Inf. Fusion 81 (2022) 84–90.

[48] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, Adv. Neural Inf. Process. Syst. 35 (2022) 507–520.
[49] Z. Jia, Controlling the overfitting of heritability in genomic selection through cross validation, Sci. Rep. 7 (1) (2017) 13678.
[50] Y. Liu, V. Gopalakrishnan, An overview and evaluation of recent machine learning imputation methods using cardiac imaging data, Data 2 (1) (2017) 8.
[51] W.C. Willett, Balancing life-style and genomics research for disease prevention, Science 296 (5568) (2002) 695–698.
[52] G.W. Miller, D.P. Jones, The nature of nurture: refining the definition of the exposome, Toxicol. Sci. 137 (1) (2014) 1–2.