ESC European Society of Cardiology

European Heart Journal - Digital Health (2023) **4**, 337–346
https://doi.org/10.1093/ehjdh/ztad033

ORIGINAL ARTICLE

A comparative study of model-centric and data-centric approaches in the development of cardiovascular disease risk prediction models in the UK Biobank

Mohammad Mamouei [1,2,*], Thomas Fisher[1,2], Shishir Rao[1,2], Yikuan Li[1,2], Ghomalreza Salimi-Khorshidi[1,2], and Kazem Rahimi[1,2,3]

[1]Deep Medicine, Oxford Martin School, University of Oxford, 1st Floor, Hayes House, 75 George Street, Oxford OX1 2BQ, UK; [2]Nuffield Department of Women's and Reproductive Health, Medical Science Division, University of Oxford, Oxford, UK; and [3]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

Received 22 November 2022; revised 1 April 2023; online publish-ahead-of-print 15 May 2023

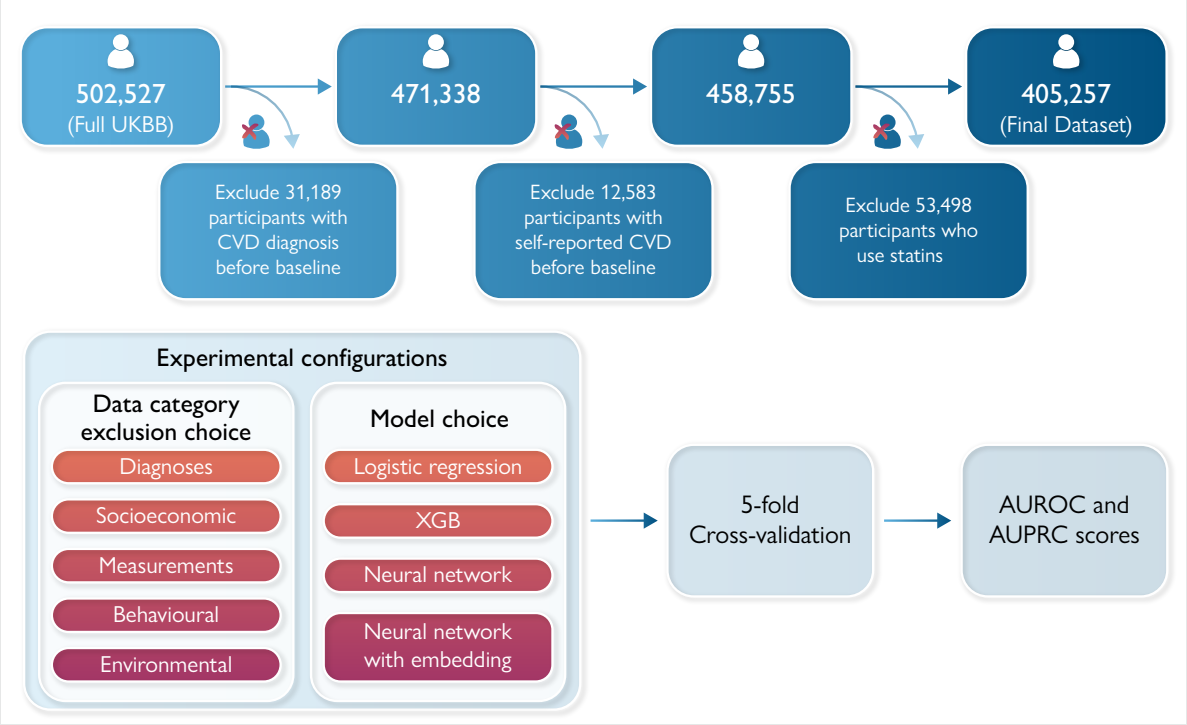
Aims A diverse set of factors influence cardiovascular diseases (CVDs), but a systematic investigation of the interplay between these determinants and the contribution of each to CVD incidence prediction is largely missing from the literature. In this study, we leverage one of the most comprehensive biobanks worldwide, the UK Biobank, to investigate the contribution of different risk factor categories to more accurate incidence predictions in the overall population, by sex, different age groups, and ethnicity.

Methods and results The investigated categories include the history of medical events, behavioural factors, socioeconomic factors, environmental factors, and measurements. We included data from a cohort of 405 257 participants aged 37–73 years and trained various machine learning and deep learning models on different subsets of risk factors to predict CVD incidence. Each of the models was trained on the complete set of predictors and subsets where each category was excluded. The results were benchmarked against QRISK3. The findings highlight that (i) leveraging a more comprehensive medical history substantially improves model performance. Relative to QRISK3, the best performing models improved the discrimination by 3.78% and improved precision by 1.80%. (ii) Both model- and data-centric approaches are necessary to improve predictive performance. The benefits of using a comprehensive history of diseases were far more pronounced when a neural sequence model, BEHRT, was used. This highlights the importance of the temporality of medical events that existing clinical risk models fail to capture. (iii) Besides the history of diseases, socioeconomic factors and measurements had small but significant independent contributions to the predictive performance.

Conclusion These findings emphasize the need for considering broad determinants and novel modelling approaches to enhance CVD incidence prediction.

* Corresponding author. Tel: +44 1865 617200, Fax: +44 1865 617202, Email: mohammad.mamouei@wrh.ox.ac.uk

Graphical Abstract



Keywords Cardiovascular disease risk modelling • Model-centric modelling • Data-centric modelling • Machine learning • Predictors of cardiovascular risk

Introduction

Cardiovascular diseases (CVDs) are associated with a broad range of risk factors including genetic, behavioural, psychological, socioeconomic, environmental factors, and the history of diseases and treatments.[1–3] The identification of individuals who have a high risk of CVDs is an effective strategy to initiate further evaluations and treatments. Therefore, CVD risk models such as QRISK3, Framingham, SCORE, and the model recommended by the American Heart Association/American College Cardiology (AHA/ACC) have become an integral part of clinical practice and research. Rooted in the statistical modelling tradition, these models are characterized by (i) a small number of predictors pertaining to well-established risk factors such as hypertension, age, smoking status, diabetes, and composite predictors such as total cholesterol:high-density lipoprotein cholesterol ratio, and (ii) simple, interpretable, functional forms. Therefore, improvements to the predictive performance of these models may be achieved using more advanced models (model-centric approach), albeit at the cost of interpretability, or by adding new predictors (data-centric approach).

Over the past few decades, machine learning (ML) has introduced a paradigm shift in predictive modelling. Owing to their improved functional forms and their ability in extracting complex patterns from high-dimensional, multimodal data, new ML models with little or no feature engineering have achieved unprecedented predictive performance across different fields. This model-centric approach has inspired much of the recent CVD risk model studies, albeit often accompanied with the inclusion of new predictors. Weng *et al.*[4] compared the performance of several ML CVD risk models with the ACC/AHA model using electronic health records (EHRs). In addition to the predictors from the ACC/AHA model, they included 22 new predictors in ML models. Compared with the ACC/AHA model, the neural network (NN) model improved area under the receiver operating characteristic (AUROC) curve by 3.6%.[4] Similar findings have been reported in other studies where ML models are shown to outperform clinical CVD risk models as well as statistical models such as Cox regression with the same predictors.[5,6]

The data-centric approach has a longer history. The Framingham study, a US cohort study with prospectively collected data over several generations, helped establish major risk factors of CVDs, namely, age, sex, high blood pressure, smoking, dyslipidaemia, and diabetes, which are used in the Framingham risk model, and many others.[7] But the study also laid the foundations for others that used primary and secondary data to discover other determinants of cardiovascular health. QRISK3 was developed from a large UK cohort using retrospective EHRs, included several new predictors such as the diagnosis of rheumatoid arthritis, chronic kidney disease (CKD), severe mental illness, and erectile dysfunction.[8] In recent years, the genetic determinants of CVDs have attracted much interest.[9] Khera *et al.*[10] and Inouye *et al.*[11] showed polygenic risk scores can identify those who have four times higher risk of CVDs. The addition of the polygenic risk score proposed by Inouye *et al.* to the established risk factors improved the discrimination of CVD risk prediction by 3.7%. Others have shown significant associations between environmental factors such as air pollution, noise, access to green space, and built environment with overall health and CVD.[12–17]

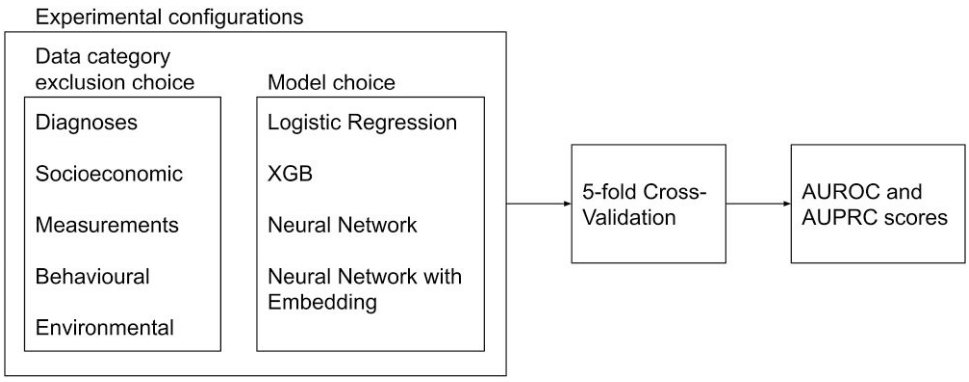


Figure 1 An outline of the study design.

The most promising approach is to align model-centric efforts with the increasing complexity of augmented data. Several studies have investigated the applications of novel deep learning (DL) models on longitudinal EHR for the prediction of CVD events with promising results.[18–22] In these studies instead of relying on clinically established risk factors, the entire history of medical events is used to train models. In doing so, novel DL models, namely sequence models, may offer clear advantages relative to conventional statistical and ML models. Novel neural sequence models can learn from the sequence of medical events while using conventional models the information pertaining to the sequence of medical events is either completely lost or requires manual feature engineering (assuming the relationship is known *a priori*). Li *et al* trained a large transformer-based model, BEHRT, on the entire patients' EHR to predict the risk of CVD events and compared the results with QRISK3, Framingham, and ASSIGN. This model substantially outperformed all the conventional models on several CVD risk prediction tasks.[23]

Despite the wealth of studies available on different determinants of CVDs, no study has provided a comparative investigation of the complementarity and contribution of different predictors in CVD risk prediction. Additionally, the interplay between the data- and model-centric approaches in the context of CVD risk prediction has not been analysed before. This study aims to fill this gap. We leveraged one of the most comprehensive biobanks worldwide, the UK Biobank, extracted information about demographic, socioeconomic, anthropometric, physiological, behavioural, environmental factors, and disease history (linked EHR and self-reported) for a cohort of 405 257 participants and analysed the independent contribution of different categories of predictors to the accuracy of predictions. Moreover, we used a variety of ML and DL models, namely, logistic regression (LR), gradient-boosted trees, and several NN models, including a transformer-based sequence model, BEHRT. The results were benchmarked against QRISK3. An outline of the study set-up is shown in *Figure 1*.

The main contributions of this study are:

(1) Delivering a better understanding of the independent contribution of different predictor categories to CVD risk prediction within a large cohort of 405 257 individuals in the UK.
(2) Evaluating the contribution of different predictor categories within the two sexes, different age groups, and ethnicities to evaluate possible discrepancies and biases.
(3) Identification of predictors associated with increased likelihood of CVD.
(4) Incorporating statistical, ML, and DL models, including the state-of-the-art sequence model, BEHRT, to provide a better understanding of the potentials and limitations of the model-centric and data-centric approaches in the context of CVD risk prediction.
(5) Using an established clinical model, QRISK3, as a benchmark to inform the expected gains.

Methods

Study design and cohort selection

This is a large population-based, cohort study using the UK Biobank baseline data and the linked in-patient hospital data for all participants ($N = 502\,527$). The cohort was recruited in the UK between 2006 and 2010. Although the UK Biobank cohort is not representative of the sampling population and there is evidence of a 'healthy volunteer' selection bias, valid assessment of exposure–disease relationships may be widely generalizable and does not require participants to be representative of the population at large, which makes it suitable for our study.[24]

We included all the participants without any CVD at the baseline based on both self-reported medical conditions and linked hospital in-patient records using the codes detailed in the Supplementary material online, *Appendix S1*. Individuals who reported the regular use of statin (simvastatin, atorvastatin fluvastatin, pravastatin, rosuvastatin) were excluded from the analysis. This led to a cohort of 405 257 participants who met the inclusion criteria (*Figure 2*). And descriptive statistics of the cohort are reported in *Table 1*.

Predictors

We included a variety of predictors in our models according to previous research on CVD and its determinants. In total, we included nine categories of predictors namely demographics, socioeconomics, anthropometric, home location, indicators of cardiac function, behavioural factors, self-reported medical conditions, medical diagnoses, and environmental factors, each category containing multiple predictors. These data were represented with 1110 continuous and binary variables. All variables in each category are listed at the end of the Supplementary material.

QRISK3

The predictors of QRISK3 were included. Predictors pertaining to diseases were ascertained based on EHR and self-reported diseases and included diabetes Type I and II, CKD, migraine, rheumatoid arthritis, systemic lupus erythematosus, severe mental illness, erectile dysfunction, and atrial fibrillation. Further details are included in Supplementary material online, *Table S1*. The use of atypical antipsychotic medication and steroid tablets was ascertained based on participants reported regular medications during the baseline interview. The list of drugs used for ascertainment is reported in Supplementary material online, *Appendix S2*. The following predictor was not available in the UK Biobank: diagnoses of angina or heart attack in a first-degree relative younger than 60 years old. Other extracted predictors are

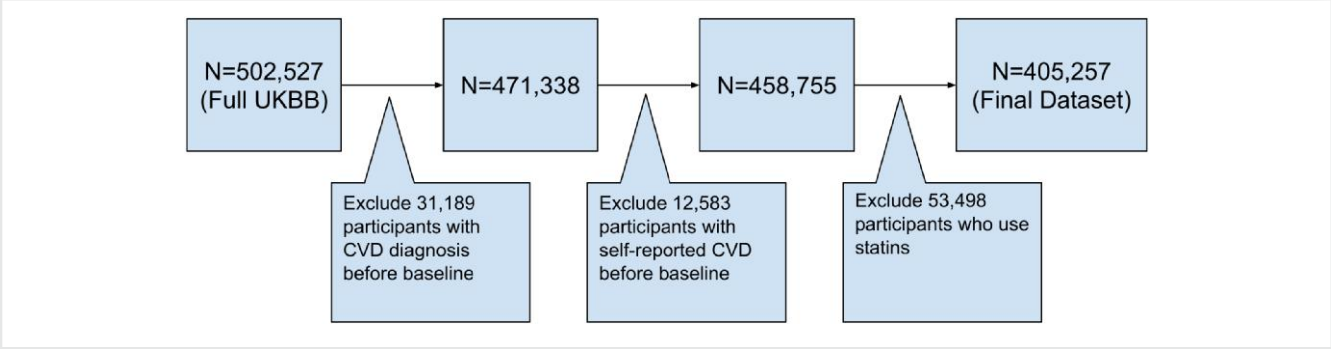


Figure 2 Schematic representation of the cohort selection.

Table 1 Overview of the cohort

Variables	Women ($n = 233\,591$)	Men ($n = 171\,666$)
Age, mean (SD)	55.57 ± 7.99[a]	55.34 ± 8.22
Townsend deprivation index, mean (SD)	−1.41 ± 3.00	−1.32 ± 3.12
Ethnicity: British (%)	87.94%	88.12%
Ethnicity: any other white background (%)	3.77%	2.88%
Ethnicity: Irish (%)	2.52%	2.82%
Ethnicity: Caribbean (%)	1.03%	0.77%
Ethnicity: Indian (%)	1.00%	1.18%
Ethnicity: other (%)	3.58%	4.02%
Household income: <18 000 GBP (%)	18.51%	15.45%
Household income: 18 000–30 999 GBP (%)	21.38%	20.73%
Household income: 31 000–51 999 GBP (%)	21.82%	24.98%
Household income: 52 000–100 000 GBP (%)	16.57%	21.68%
Household income: other (%)	20.61%	15.97%
Employment: in paid employment or self-employed (%)	58.57%	66.85%
Employment: retired (%)	31.08%	25.43%
Employment: other (%)	10.19%	7.52%
Qualifications: O levels/GCSEs or equivalent (%)	23.26%	18.60%
Qualifications: A levels/AS levels or equivalent (%)	12.15%	10.63%
Qualifications: College or University degree (%)	32.64%	35.64%
Qualifications: other (%)	31.01%	34.15%
BMI (% non-missing values)	27 ± 5.01 (99.48%)	27 ± 4.04 (99.33%)
Pulse rate (% non-missing values)	70 ± 10.40 (93.96%)	68 ± 11.54 (93.95%)
Systolic blood pressure (% non-missing values)	134 ± 19.19 (93.96%)	140 ± 17.36 (93.95%)
Diastolic blood pressure (% non-missing values)	81 ± 10.02 (93.96%)	84 ± 10.00 (93.95%)
Cholesterol:HDL cholesterol ratio (% non-missing values)	4 ± 1.01 (84.61%)	5 ± 1.15 (86.63%)
Age at first CVD diagnosis	64.74 ± 7.29	64.12 ± 7.38
Number of self-reported medical conditions	1.54 IQR: [1, 2]	1.29 IQR: [1, 2]
Number of diagnosed medical conditions	2.53 IQR: [0, 4]	1.83 IQR: [0, 3]
Years of follow-up	8.07 ± 1.29	7.91 ± 1.57
Number of new CVD events	12 659 (5.42%)	16 727 (9.74%)
CVD incidence rate, per 1000 person-years	7	12

[a]The values following the symbol ± show the standard deviation.

age, sex, ethnicity, smoking status, average systolic blood pressure (SBP), body mass index (BMI), cholesterol/HDL ratio.

To evaluate the predictive value of different determinants of CVD, we grouped all predictors into five categories.

Socioeconomic category

We included the demographic and socioeconomic predictors of sex and ethnic background as demographic factors and household income before tax, current employment status, Townsend deprivation index at recruitment, and education qualifications as socioeconomic factors.

Measurements category

We included common anthropometric risk factors such as BMI and birth weight and well-established risk factors for CVD outcomes such as average diastolic blood pressure, SBP and pulse rate, and cholesterol to HDL ratio.

Behavioural category

Smoking status and alcohol consumption status were included in the models.

Diagnosis category

This category consists of two different sources of medical records that are available in the UKBB.

(1) Self-reported medical conditions: The baseline medical conditions were assessed for all the participants and were classified into 445 different categories. Since each participant may have more than one present medical condition, we used one-hot encoding to represent the baseline medical conditions: each participant was represented as a binary vector with 445 elements where each element stands for one medical condition with present conditions coded as 1 and otherwise 0.
(2) Diagnosed medical conditions: We included the entire medical history, as recorded by ICD-10 codes, in the models. ICD-10 is a hierarchical coding system; at the highest level, it categorizes diseases into 22 chapters. Operating at this level leads to a loss of specificity as distinct diseases fall within the same chapter. The lowest level of the coding hierarchy captures the most detailed description of the diagnoses, but there are 19 155 distinct ICD-10 codes in the UK Biobank and the great majority of these codes appear only a few times each. Operating at this level leads to many features, most of which do not have enough training examples to learn from. To address this, we used a rule-based method to map the diagnoses to higher levels if the numbers of events at the lowest level of the hierarchy were insufficient. Here, we considered events with fewer than 1000 occurrences insufficient, leading to 595 binary columns, some at the ICD-10 subchapter level, and some at lowest level of the hierarchy ICD-10 4-digit codes.

Environmental category

We included the annual average concentration of $PM_{2.5}$ (particulate matter with an aerodynamic diameter of <2.5 μm), PM_{10} (particulate matter with diameter ≤10 μm), PM_{coarse} (particulate matter with an aerodynamic diameter between 2.5 and 10 μm), $PM_{2.5}$ absorbance (a measurement of the blackness of $PM_{2.5}$ filters—a proxy for elemental carbon, which is the dominant light absorbing substance), NO_2 (nitrogen dioxide), and NO_x (nitrogen oxides) which were calculated using a Land Use Regression model developed by the ESCAPE project.[25,26] We included the average values of NO_2 and PM_{10} concentration data for 2010. Traffic-related predictors, namely vicinity to major roads, inverse distance to the nearest major road, inverse distance to the nearest road, sum of road length of major roads within 100 m, total traffic load on major roads, traffic intensity on the nearest major road, and traffic intensity on the nearest road were also calculated in the ESCAPE project and were included in the analysis.

Data pre-processing

The continuous predictors were normalized to aid with algorithm convergence and categorical columns were converted into one-hot encoding vectors. The missing data were imputed using Multiple Imputation by Chained Equations with gradient-boosted trees (miceforest 5.4.0 package in Python). We used four iterations and during each, a random subsample containing half of the observations was used for imputation. Comparing the distribution of the imputed data and non-imputed data verified that they were aligned.

Outcome and time window

The outcome of interest in this study was the incidence of CVD, including coronary heart disease, myocardial infarction, heart failure, and valvular heart disease. All outcomes were ascertained using ICD codes (see Supplementary material online, *Appendix S2*) from the linked in-patient hospital data after baseline. Consistent with established risk models, we considered a 10-year follow-up period.

Model development and evaluation

To examine the independent contribution of each category we excluded each from the total collection of 1100 predictors, models were trained on the reduced subset and evaluated using five-fold cross-validation. We used LR and extreme gradient-boosted trees (XGB). These models have been reported to have superior performance relative to clinical CVD risk models.[4] We also included a multi-layer feedforward NN and a similar network but with the addition of an embedding layer for the categorical predictors (NN-EMB).

We also trained a sequence model, BEHRT, on the sequence of medical diagnoses and compared its performance with other models to investigate whether capturing the sequence of medical events, rather than considering their absence or presence alone (known as the bag-of-words representation), can improve the predictive performance. For this comparison, all individuals without any medical diagnoses were excluded from the analysis, leading to a separate cohort of 234 938 individuals. Other than a higher CVD incidence rate in this cohort, other characteristics were comparable in the two cohorts (see Supplementary material online, *Table S2*). While it is preferable to limit the analysis to individuals with much longer sequences of medical events, this will lead to very small subgroups and due to the absence of sufficient training examples BEHRT-like models perform poorly.[27,28] Details about the BEHRT can be found in the related publication.[21]

To benchmark the results, we also included QRISK3 in our analysis, a CVD risk model developed and validated in the UK population.[29] This model was implemented using Cox regression, a survival model which accounts for censoring. We would like to highlight that we used a different definition of CVD than that of QRISK3, for instance, we included valvular heart disease as an outcome and angina as a predictor. Similar to other models, QRISK 3 was trained and evaluated using cross-validation. The hyperparameters of all models were selected using Bayesian search and are reported in Supplementary material online, *Appendix S3* and *Table S3*.

For the classification models, censored participants were considered event-free. This is a limitation arising from the comparison of classification models with survival models. It is known to produce bias and leads to underestimation of risk in classification models, but compared with the alternative of excluding all censored patients, it is the preferred choice.[30] Lastly, one of the main objectives of our study is to investigate the extent to which different data modalities contribute to the prediction of CVD incidence. Since no subset of data is expected to systematically affect the censoring of patients, we believe the comparison remains valid.

We used the AUROC and the area under the precision recall curve (AUPRC) (or average precision) as metrics to evaluate the performance of our models during cross-validation.

Results

Detailed results from five-fold cross-validation for all models and all data exclusions are reported in *Table 2*, which highlights the best AUROC and AUPRC scores in each column in bold text. We refer to the data sets by which category of predictors has been excluded. For example, -SocioEcon refers to the data set where all of the socioeconomic predictors are removed. An extended version of this table that includes the standard deviation of the models is included in Supplementary material online, *Table S4*. The ROC, precision–recall,

Table 2 Area under the receiver operating characteristic and area under the precision recall curve for all models with different subsets of predictors

Model	Model performance for different predictor exclusions (%AUROC|%AUPRC)					
	Diagnoses	SocioEcon	Measurements	Behavioural	Environment	All
LR	71.55|16.18	72.79|17.72	72.93|**17.90**	73.01|**17.89**	73.20|18.02	73.21|**18.03**
XGB	**71.75**|**16.38**	72.85|**17.73**	72.81|17.76	72.98|17.84	73.24|**18.18**	73.20|17.99
NN	71.58|16.06	72.81|17.52	72.85|17.62	72.96|17.63	73.19|17.83	73.19|17.82
NN-EMB	71.64|15.97	**72.91**|17.60	**72.99**|17.69	**73.09**|17.72	**73.26**|17.84	**73.28**|17.89
QRISK3	69.50|16.38					

Each column highlights the performance of models after excluding a subset of predictors.
Bold figures represent the best measure of performance (AUROC/AUPRC) in each column.

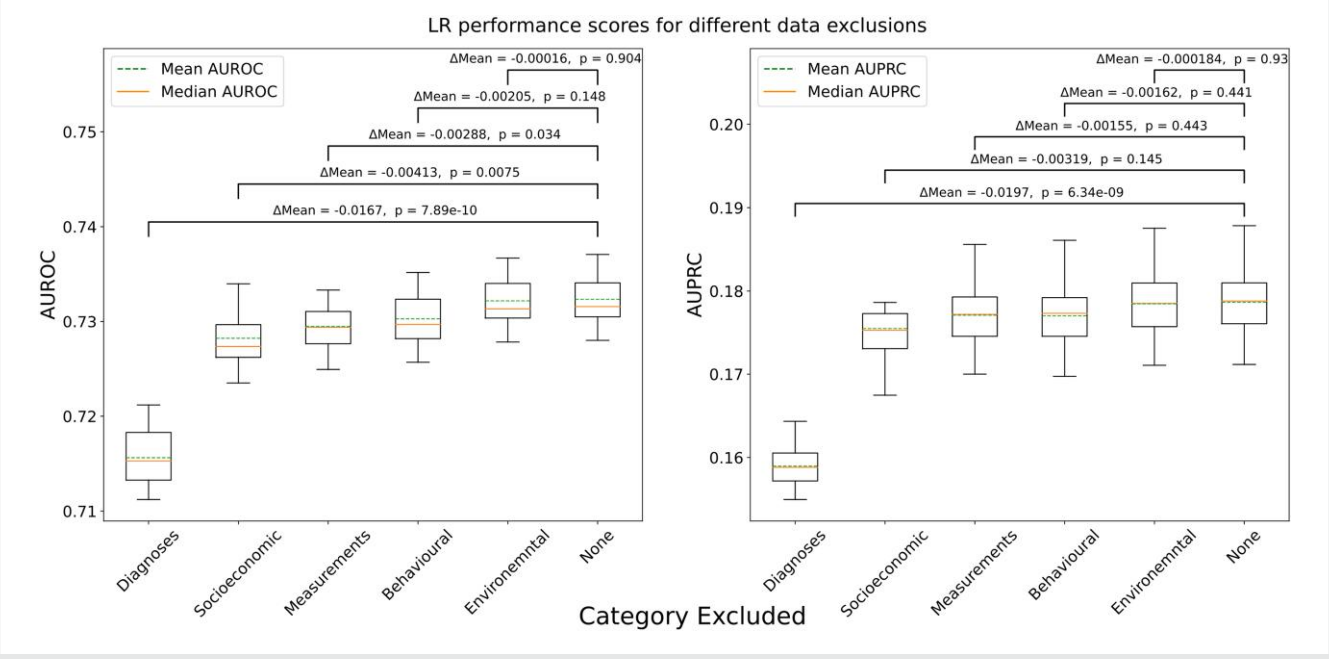


Figure 3 The performance of the logistic regression model with different data exclusions obtained with five-fold cross-validation: area under the receiver operating characteristic (left plot) and area under the precision recall curve (right plot). The pairwise mean differences are noted above the horizontal lines. The *P*-values are obtained using Welch's *t*-tests.

and calibration curves for all models are presented in Supplementary material online, *Figure S1*. The average AUROC and AUPRC of QRISK3, without any changes to its predictors, was similarly obtained from five-fold cross-validation and is included in the table for comparison.

The differences in performance arising from the use of different models are small. However, regarding average precision, XGB and LR deliver better performances relative to the NN models. The NN with learnt embeddings (disease representations), i.e. NN-EMB, obtains the best discrimination in all cases where the extensive medical history is used as predictors. Compared with the NN model, this model delivers around 0.1% improvement in discrimination as well as a robust improvement in precision. This clearly highlights the value of learned representations. It is also important to note that all models that use the entire available medical information substantially outperform the QRISK3 model.

To facilitate the interpretation of the differences arising from data exclusions, we evaluated the statistical significance of the differences in AUROC and AUPRC using Welch's *t*-test. Since, the patterns remain largely true for other models, for brevity only the results for the LR model are visualized in *Figure 3*.

The exclusion of the history of diagnoses has the largest effect on the performance of the model, resulting in a −2.97% change in AUPRC and −1.67% change in AUROC. This is followed by the socioeconomic, measurements, and behavioural categories. The exclusion of the socioeconomic predictors changed the AUPRC by −0.32% (AUROC change of 0.41%). For measurements and behavioural predictors, the exclusion resulted in −0.16% reduction in AUPRC (−0.03 and −0.02% reduction in AUROC, respectively). The environmental category despite the large number and the diversity of the included predictors made the smallest contribution to the predictive performance.

Model performance in subgroups

The predictive value of each category could be different across different subgroups. For instance, the history of medical events might be longer in older age groups and its exclusion might lead to a larger reduction in predictive performance. To investigate this, we analysed the performance of the models in different subgroups, specifically in different age groups, sexes, and ethnic backgrounds. To ensure the performance for different age subgroups is not affected by the number of observations, they were grouped based on age quartiles. For ethnicities, we used the higher-level classification used in the UK Biobank as detailed in Supplementary material online, *Appendix S4*. The findings were consistent across all models. Therefore, for brevity only the results pertaining to the LR models are presented. Apart from the diagnoses category, the exclusion of other categories did not produce notable changes in the performance of the models across different subgroups. *Figure 4* shows a summary of the results. All results pertaining to the LR model are reported in Supplementary material online, *Appendix S4*.

The analysis highlights inconsistencies both in the performance of the models and in the predictive value of medical history within various strata. Firstly, the AUPRC is notably higher for females. The value of medical diagnoses also seems to be higher for this subgroup. The higher number of self-reported and diagnosed conditions in females may provide an explanation for this (*Table 1*).

Secondly, medical history notably has a higher predictive value for the second age quartile, i.e. the subgroup between 56 and 62 years of age. This is counter-intuitive as the oldest subgroup has both larger number of diagnosed conditions and higher event rates (see Supplementary material online, *Table S5*). The lengthier medical history and more positive labels should both contribute to more precision. Three hypotheses could provide possible explanations for this observation: (i) the process of ageing accompanies more rapid changes in the medical conditions of the oldest age group and that could make a 10-year risk prediction more challenging. (ii) In the light of the improving quality of care over time,[31,32] the oldest subgroup may have experienced a lower quality of care and therefore might have less informative medical history, and finally (iii) competing risks that increasingly become more important in the oldest age subgroup might be less adequately captured in medical diagnoses alone.

Regarding differences within various ethnicities, the substantial differences in the number of participants in each stratum and the broad standard deviations make the interpretation of findings challenging; however, the Chinese and the mixed subgroup were least affected by the exclusion of medical history. The analysis of the number of self-reported and diagnosed medical condition shows that the Chinese subgroup have substantially less recorded conditions (see Supplementary material online, *Table S6*). The Chinese had the average 1.00 self-reported and 1.63 diagnosed conditions compared with the average of 1.44 self-reported and 2.23 diagnosed conditions in the rest of the population. We could not identify plausible explanations for the 'other' subgroup. The Asian and the black subgroups had the most notable decreases in AUROC and AUPRC after the exclusion of the medical diagnoses.

Modelling the sequence of diagnoses

Table 3 shows the performance of BEHRT with medical diagnoses only. This analysis was carried out on the cohort of patients with at least one diagnosis (see Supplementary material online, *Table S2*). For comparison, we similarly retrained the LR model with only medical diagnosis. As both models were trained on only medical diagnosis, the differences in their performance could be attributed to architectural merits of BEHRT and the importance of capturing the sequence of medical events. QRISK, without any changes to its predictors, was also retrained on the same cohort and included in *Table 3*. The AUROC and AUPRC are averaged over five-fold cross-validation.

Using medical diagnoses alone, BEHRT significantly improves the performance of the LR model. This highlights the importance of contextualized embeddings and the sequence of medical events. The model also achieves higher discrimination and precision compared with QRISK3.

Discussion

Over the past few decades, numerous studies have shown the association of socioeconomic, physiological, behavioural, and environmental factors with CVDs. Additionally, many studies have shown that clinical CVD risk models can be improved by incorporation of new predictors, use of new models, or a combination of the two. However, joint investigation of both model-centric and data-centric approaches is scarce, making it challenging to draw clear conclusions about the merits, limitations, and synergies of the two.

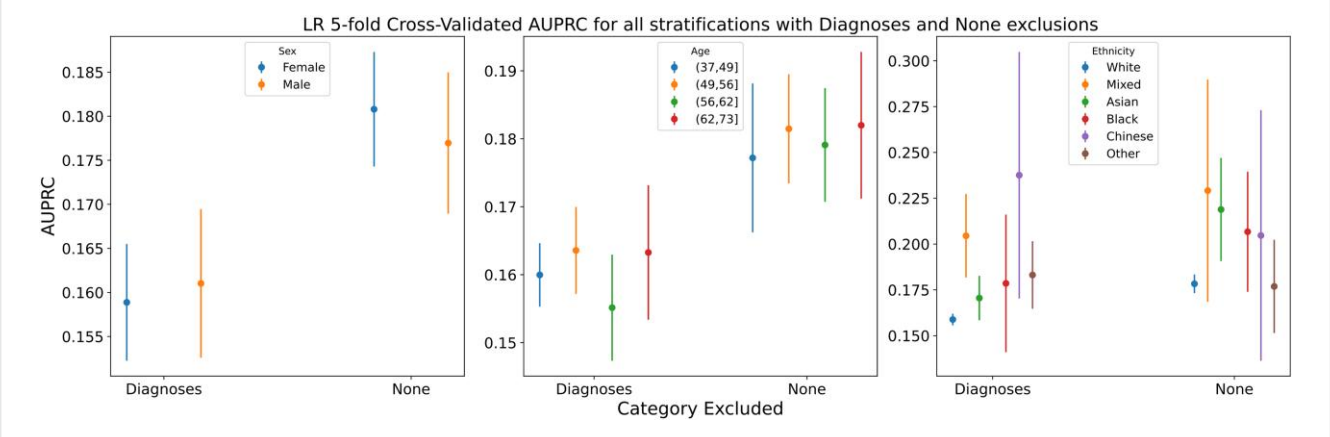


Figure 4 The stratified analysis of the area under the precision recall curve for the logistic regression model with and without the diagnoses category: sex-stratified analysis (left plot), age-stratified analysis (middle plot), and ethnicity-stratified analysis (right plot). The error bars show the standard deviation of the metrics within five-fold cross-validation.

Table 3 Area under the receiver operating characteristic and area under the precision recall curve of BEHRT and logistic regression trained on only diagnosis data in the cohort of participants with at least one diagnosis

Model	Model performance (%AUROC|%AUPRC)
LR	65.07|15.54
BEHRT	**69.83|17.90**
QRISK	69.50|16.38

Bold figures represent the best measure of performance (AUROC/AUPRC) in each column.

In this study, we investigated the independent contribution of various CVD predictor categories to the performance of risk models in a cohort of 405 257 UK Biobank participants. This was complemented by a model-centric approach and incorporation of several statistical and DL models in the analysis. The results highlight that using a more comprehensive history of diseases instead of established CVD risk factors which are used in existing clinical risk models could substantially improve the identification of individuals who are at a high risk of CVD incidence. In our study, we were limited to the use of self-reported and in-patient diagnosis data in the UK Biobank. The incorporation of other clinical data such as operations, lab tests, measurements, and medical records from general practitioners could lead to even more substantial improvements in CVD risk prediction.[21,23,27]

While in our study socioeconomic, measurements, behavioural, and environmental predictors independently contributed little to the accuracy of predictions, in the absence of any causal assumptions, this does not reflect a lack of strong causal pathways between these predictors and CVD incidence.

We analysed the coefficients of the LR model with *P*-value smaller than 0.001. In the model with the diagnosis category excluded, inability to work due to sickness, age, and being male were associated with a higher CVD risk. The remaining significant predictors that increased the risk of CVD were BMI, SBP, standing height, cholesterol/HDL ratio, inverse distance to nearest major road, $PM_{2.5}$, traffic intensity on nearest major road, nitrogen oxides, Townsend deprivation index, and pulse rate. In the models that included the diagnosis category, the positive and negative values were largely similar. The top 10 features that increased the risk of CVD were Wolff–Parkinson–White syndrome, aortic aneurysm and dissection, palpitations, sickle cell disease, pericardial problem hyperprolactinaemia, condition originating in the perinatal period, non-Hodgkin's lymphoma, Sjogren's syndrome/Sicca syndrome, and systemic sclerosis. The features that reduced the likelihood CVD were less informative and many of them were established risk factors of mortality. For instance, HIV had the highest negative coefficient. This is an artefact of using a classification model and marking censored patients as event-free. Feature importance plots for these models are included in Supplementary material online, *Appendix S5*.

We observed substantial inconsistencies in the performance of models within the two sexes, age groups, and ethnicities. Individuals of Chinese background and males were observed to have less reported and diagnosed conditions. While the findings should be interpreted in the context of the UK Biobank and possible biases in its recruitment process, it might reflect a broader difference in how these subgroups interact with the healthcare system. The use of other observational data such as the Clinical Practice Research Datalink (CPRD) could shed light on this. Two recent studies based on CPRD show that compared with other ethnic groups, people of Chinese ethnicity have the lowest mean number of EHR-determined long-term conditions and lowest prevalence of complex multimorbidity across all age groups.[33,34] Our study suggests the predictive value of diagnosed and self-reported conditions within the Chinese ethnicity is lower than others ethnic groups. A possible explanation to this might be lower diagnoses of medical conditions in this subpopulation. It is worth mentioning that people of Chinese ethnicity are reported to have the best mortality outcome and health-related quality of life compared with other ethnic groups; however, this may be attributed to their substantially better self-care.[33,35] Viewed together, this highlights a possible mechanism for bias in similar risk models.

Comparing the performance of different models shows that using tabular representation of data, an interpretable, statistical model namely, LR, can deliver comparable performance with, and outperform, ML and DL alternatives. But we showed that a neural sequence model trained on the sequence of diagnoses can substantially improve the performance of a LR model that was trained on the tabular representation of the same data. The improved predictive performance of such models should be viewed along with the challenges pertaining to their interpretability.

Compared with QRISK3, all other models were miscalibrated. This is a limitation; however, it should not be immediately viewed as an indication of poor predictive performance. QRISK3, although well calibrated, generally produced much lower risk scores relative to other models. In the study cohort, the highest decile of risk based on QRISK3 had an average predicted risk score of 0.26 from which 21% had incident CVDs. The ML models produced much broader range of risk estimates. The highest decile of risk in the ML models was around 0.7 from which nearly 30% had incident CVDs. While this is a notable overestimation of absolute risk, it is important to note that the ML models produced much more graded risk scores across the population and the risk estimates positively correlated with the likelihood of CVDs; as evident from the monotonically increasing calibration curve. The higher resolution of the ML model translates into better discrimination and precision, i.e. better ability of the models in distinguishing higher risk individuals from lower risk individuals in a classification setting, as seen in AUROC and AUPRC. Contrary to this, QRISK3 groups many individuals who may have different underlying risks in the same risk brackets. In the light of this, the estimates derived from miscalibrated ML models should be viewed as a measure of rank rather than the absolute risk of CVD incidence. The risk thresholds for binary classification are also commonly selected based on the discrimination/precision trade-offs and not the default 0.5. In practical applications where predicted risk scores of uncalibrated models may be incorrectly interpreted as the likelihood of disease incidence, *post hoc* calibration could deliver accurate absolute risk estimates at no or minimal cost to predictive performance.[36] These statements do not undermine the importance of calibration curves, but rather highlight the necessity of considering precision, discrimination, the shape and range of the calibration curve, as well as its alignment with the diagonal line.

In addition, to training the BEHRT model on the sequence of diagnoses, inspired by Targeted-BEHRT, we used an extended version of BEHRT, to combine the static predictors with the sequence of medical diagnoses.[37] We explored several early and joint fusion models to combine the static and longitudinal data. All models delivered lower performance than the LR model with all predictors. This underlines the need for tailored, more data-efficient models for fusion of various data modalities.

This study was first and foremost a methodological investigation of the predictive value of various predictors and modelling approaches; without further external validations, considerations pertaining to model explainability and other practical implications such as the availability of data, the developed models *per se* are not intended for clinical use case.

Limitations

Our analysis was based on self-reported and in-patient data available in the UK Biobank. The findings should be viewed under this limitation.

The predictors of CVDs are numerous and many of them are available in the UK Biobank. We have only analysed a relatively small subset of these predictors. Genetic, physical activity, sleep quality, electrocardiogram measurements, dietary data, and medications, to name a few, are all important determinants that we did not analyse in our study. Given the importance of CVD risk prediction in targeting risk-mitigating interventions, we hope similar studies can be conducted in the future to shed light on the predictive value of other categories in comparison with established risk factors, such as the ones used in QRISK3, and EHR-derived predictors that have become an active area of research in recent years. The authors believe more research in this area could facilitate rapid improvements of existing risk models.

In the absence of other data sources with the same diversity of predictors as the UK Biobank, we could not externally validate the models. The generalizability of the findings should be viewed under this limitation.

Ascertaining the exact time of CVD incidence based on incomplete EHR is not possible. We cannot rule out that some of the recorded conditions such as pericarditis may have been complications of a previously diagnosed CVD that was not recorded in hospital in-patient EHR until after the index time. Therefore, clinical context and caution should be used in the interpretation of results.

Author contributions

M.M. and K.R. conceived the idea for this study. M.M., T.F., and K.R. contributed to the study design. S.R. and Y.L. advised on the development and implementation of BEHRT. M.M. and T.F. performed the data preprocessing and model development. T.F. reported the results. M.M., T.F., K.R., and G.S.-K. contributed to the interpretation of results. M.M. and T.F. wrote the first draft. All authors contributed to revisions of the manuscript and approved the final version. All authors had full access to the data.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health* online.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. Their input has immensely enriched the work, especially the discussion and limitations sections of the article.

Funding

This work is supported by the PEAK Urban programme, funded by UKRI's Global Challenge Research Fund (ES/P011055/1). The following authors are supported by grants from the British Heart Foundation (BHF): Y.L. and K.R. (FS/PhD/21/29110) and K.R. (PG/18/65/33872); K.R. is also in receipt of funding from Oxford NIHR Biomedical Research Centre and the Oxford Martin School (OMS), University of Oxford. M.M. is in receipt of funding from Novo Nordisk (CRR00515 HE00.01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and not necessarily those of the OMS, the BHF, the UKRI, the NIHR, or Novo Nordisk.

Conflict of interest: None declared.

Data availability

The data used in the study are available to approved researchers via the UK Biobank.

References

1. Anene-Nzelu CG, Lee MCJ, Tan WLW, Dashi A, Foo RSY. Genomic enhancers in cardiac development and disease. *Nat Rev Cardiol* 2022;**19**:7–25.
2. Yusuf S, Joseph P, Rangarajan S, Islam S, Mente A, Hystad P, *et al.* Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet* 2020;**395**:795–808.
3. Chaulin AM, Duplyakov DV. Environmental factors and cardiovascular diseases. *Gig i Sanit* 2021;**100**:223–228.
4. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;**12**:e0174944.
5. Cho SY, Kim SH, Kang SH, Lee KJ, Choi D, Kang S, *et al.* Pre-existing and machine learning-based models for cardiovascular risk prediction. *Sci Rep* 2021;**11**:8886.
6. Alaa AM, Bolton T, di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One* 2019;**14**:e0213653.
7. D'Agostino RB, 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**:743–753.
8. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;**357**:j2099.
9. Knowles JW, Ashley EA. Cardiovascular disease: the rise of the genetic risk score. *PLoS Med* 2018;**15**:e1002546.
10. Khera A, Chaffin M, Aragam K, Emdin C, Klarin D, Haas M, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**:1219–1224.
11. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, *et al.* Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J Am Coll Cardiol* 2018;**72**:1883–1893.
12. Greenfield BK, Rajan J, McKone TE. A multivariate analysis of CalEnviroScreen: comparing environmental and socioeconomic stressors versus chronic disease. *Environ Health* 2017;**16**:131.
13. Floud S, Blangiardo M, Clark C, de Hoogh K, Babisch W, Houthuijs D, *et al.* Exposure to aircraft and road traffic noise and associations with heart disease and stroke in six European countries: a cross-sectional study. *Environ Health* 2013;**12**:89.
14. Thacher JD, Hvidtfeldt UA, Poulsen AH, Raaschou-Nielsen O, Ketzel M, Brandt J, *et al.* Long-term residential road traffic noise and mortality in a Danish cohort. *Environ Res* 2020;**187**:109633.
15. Vienneau D, Schindler C, Perez L, Probst-Hensch N, Röösli M. The relationship between transportation noise exposure and ischemic heart disease: a meta-analysis. *Environ Res* 2015;**138**:372–380.
16. Bhatnagar A. Environmental determinants of cardiovascular disease. *Circ Res* 2017;**121**:162–180.
17. Mamouei M, Zhu Y, Nazarzadeh M, Hassaine A, Salimi-Khorshidi G, Cai Y, *et al.* Investigating the association of environmental exposures and all-cause mortality in the UK Biobank using sparse principal component analysis. *Sci Rep* 2022;**12**:9239.
18. Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. *Adv Neural Inform Process Syst* 2016:3504–3512.
19. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016;**6**:26094.
20. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: a convolutional net for medical records. *IEEE J Biomed Health Inform* 2017;**21**:22–30.
21. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, *et al.* BEHRT: transformer for electronic health records. *Sci Rep* 2020;**10**:7155.
22. Solares A Jr, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, *et al.* Deep learning for electronic health records: a comparative review of multiple deep neural architectures. *J Biomed Inform* 2020;**101**:103337.
23. Li Y, Salimi-Khorshidi G, Rao S, Canoy D, Hassaine A, Lukasiewicz T, *et al.* Validation of risk prediction models applied to longitudinal electronic health record data for the prediction of major cardiovascular events in the presence of data shifts. *Eur Heart J - Digit Health* 2022;**4**:535–547.
24. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol* 2017;**186**:1026–1034.
25. Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X, *et al.* Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—the ESCAPE project. *Atmos Environ* 2013;**72**:10–23.

26. Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, *et al.* Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; results of the ESCAPE project. *Environ Sci Technol* 2012;**46**: 11195–11205.
27. Li Y, Mamouei M, Salimi-Khorshidi G, Rao S, Hassaine A, Canoy D, *et al.* Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *arXiv*. 2021.
28. Hestness J, Narang S, Ardalani N, Diamos G, Jun H, Kianinejad H, *et al.* Deep learning scaling is predictable, empirically. *arXiv* 2017:1712.00409.
29. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009; **339**:b2584.
30. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ* 2020;**371**:m3919.
31. Bhatnagar P, Wickramasinghe K, Wilkins E, Townsend N. Trends in the epidemiology of cardiovascular disease in the UK. *Heart* 2016;**102**:1945–1952.
32. Conrad N, Judge A, Tran J, Mohseni H, Hedgecott D, Crespillo AP, *et al.* Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet* 2018;**391**:572–580.
33. Stafford M, Knight H, Hughes J, Alarilla A, Mondor L, Pefoyo Kone A, *et al.* Associations between multiple long-term conditions and mortality in diverse ethnic groups. *PLoS One* 2022;**17**:e0266418.
34. Hayanga B, Stafford M, Saunders CL, Bécares L. Ethnic inequalities in age-related patterns of multiple long-term conditions in England: analysis of primary care and nationally representative survey data. *medRxiv*. 2022. doi: 10.1101/2022.08.05.22278462.
35. Watkinson RE, Sutton M, Turner AJ. Ethnic inequalities in health-related quality of life among older adults in England: secondary analysis of a national cross-sectional survey. *Lancet Public Health* 2021;**6**:e145–e154.
36. Rajaraman S, Ganesan P, Antani S. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLoS One* 2022;**17**:e0262838.
37. Rao S, Mamouei M, Salimi-Khorshidi G, Li Y, Ramakrishnan R, Hassaine A, *et al.* Targeted-BEHRT: deep learning for observational causal inference on longitudinal electronic health records. *IEEE Trans Neural Netw Learn Syst* 2022:1–12.