scientific reports

OPEN

Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study

Iqbal Madakkatel[1,2]✉, Ang Zhou[1,3], Mark D. McDonnell[2,4] & Elina Hyppönen[1,3,4]✉

We present a simple and efficient hypothesis-free machine learning pipeline for risk factor discovery that accounts for non-linearity and interaction in large biomedical databases with minimal variable pre-processing. In this study, mortality models were built using gradient boosting decision trees (GBDT) and important predictors were identified using a Shapley values-based feature attribution method, SHAP values. Cox models controlled for false discovery rate were used for confounder adjustment, interpretability, and further validation. The pipeline was tested using information from 502,506 UK Biobank participants, aged 37–73 years at recruitment and followed over seven years for mortality registrations. From the 11,639 predictors included in GBDT, 193 potential risk factors had SHAP values ≥ 0.05, passed the correlation test, and were selected for further modelling. Of the total variable importance summed up, 60% was directly health related, and baseline characteristics, sociodemographics, and lifestyle factors each contributed about 10%. Cox models adjusted for baseline characteristics, showed evidence for an association with mortality for 166 out of the 193 predictors. These included mostly well-known risk factors (e.g., age, sex, ethnicity, education, material deprivation, smoking, physical activity, self-rated health, BMI, and many disease outcomes). For 19 predictors we saw evidence for an association in the unadjusted but not adjusted analyses, suggesting bias by confounding. Our GBDT-SHAP pipeline was able to identify relevant predictors 'hidden' within thousands of variables, providing an efficient and pragmatic solution for the first stage of hypothesis free risk factor identification.

Cohort studies and biobanks available for medical research are growing, both in the number of individuals included and the density of information available for the participants. These large databases hold enormous potential for innovation and provide exciting prospects for hypothesis free risk factor discovery. However, in practice, many research projects use only a set of handpicked predictors for their analyses, due to various limitations. Indeed, traditional epidemiological approaches, such as logistic regression and Cox regression are limited in number of independent variables that can be practically included in a single model. They also require lack of multicollinearity among independent variables, and without careful modelling, by default associations are assumed to be linear with no interactions between the explanatory variables. Further challenges in the multivariate context arise from the treatment of and biases caused by missing information.

Machine learning (ML), "the study of computer algorithms that allow computer programs to automatically improve through experience"[1], provides some attractive solutions for many of these challenges, and they have been found to be effective in developing predictive models based on large sets of variables. Supervised ML methods, involving labeled data, can capture complex interactions and non-linear associations among explanatory variables[2,3], often resulting in good model performance when subsequently applied to real-world data. There has been great interest in comparing model performance among different ML algorithms[4–7]. ML approaches, such as gradient boosting decision trees (GBDT)[8], support vector machines[9], K-nearest neighbors[10], and artificial

[1]Australian Centre for Precision Health, UniSA Clinical and Health Sciences, University of South Australia, Adelaide, Australia. [2]Computational Learning Systems Laboratory, UniSA STEM, University of South Australia, Mawson Lakes, Australia. [3]South Australian Health and Medical Research Institute (SAHMRI) Level 8, GPO Box 2471, Adelaide, SA 5001, Australia. [4]These authors jointly supervised this work: Mark D. McDonnell and Elina Hyppönen
✉email: iqbal.madakkatel@unisa.edu.au; Elina.Hypponen@unisa.edu.au

neural networks[11] have been found to outperform traditional risk scoring systems[4,5,12,13]. Supplementary Note online describes these ML approaches. Among the strongest approaches is GBDT, which according to a review comparing 13 different state-of-art ML methods, was ranked as the best of all methods in tasks related to predictive analytics (appreciating that no single algorithm performs the best across all datasets)[14].

In this study, our intention is not to build competing predictive models or to argue that ML methods (specifically, GBDT) is 'better', or epidemiological analyses are 'worse' but to present an approach which can accommodate their distinct strengths and limitations. We test the proof of principle of such an approach for its ability to discover potential risk factors amongst thousands of predictors by combining GBDT modelling with standard epidemiological practices. We use data from over 11,000 predictors and mortality for over 500,000 participants in the UK Biobank. Our novel analysis pipeline uses GBDT, CatBoost implementation[15] for its inherent capability to handle missing values and a large volume of data, without having to convert variables to any specific format. We screen for potentially interesting predictors using SHAP (SHapley Additive exPlanation) values[16,17], a Shapley value based additive feature attribution method, reflecting variable 'importance' to guide the selection of mortality predictors for further epidemiological modelling. Furthermore, we use penalized (LASSO) logistic regression as an alternative baseline approach and also include comparisons with another feature selection method (XGBoost[18] using five different built-in ways of calculating feature importance, as done by other studies[19–22]), and as we describe in this paper, our data suggests that the proposed GBDT-SHAP pipeline has certain advantages over them.

Methods

Participants.

The UK Biobank is a cohort of over 500,000 participants recruited between March 13, 2006 and October 1, 2010 through 22 assessment centers across England, Wales, and Scotland[23]. Data collection during the baseline assessment covered touch screen questionnaire surveys, face-to-face interviews, and physical measurements, with blood sampling and urine collection for genetic assays and biomarker assessments. Further information on disease outcomes was obtained through record linkages, including mortality statistics from the UK Office of National Statistics, cancer registrations, and hospital episodes statistics.

The outcome variable indicating the mortality status of the participants as of March 1, 2016, was created using the UK Biobank date of death field 40,000. In this study, we considered those information that were collected at the baseline assessment, including data obtained using the touchscreen questionnaires and results from clinical examinations. In addition, we included disease codings derived from linkage to cancer registrations and hospital episodes statistics. We removed baseline variables which were recorded for less than 95% of the participants. Information obtained from online follow-up surveys or sub-samples of the cohort were excluded from our analyses due to their low coverage. Supplementary Table S1 online lists all the variables included.

The UK Biobank project was approved by the National Information Governance Board for Health and Social Care and North West Multi-center Research Ethics Committee (11/NW/0382). Participants provided electronic informed consent to use their anonymized data and samples for health-related research, to be recontacted for further sub-studies and for the UK Biobank to access their health-related records[24]. This study was conducted under application number 20175 to the UK Biobank and all methods were performed in accordance with the relevant guidelines and regulations.

Model development pipeline and statistical analyses.

The GBDT-SHAP pipeline is shown in Fig. 1. As the data were not sufficiently structured for our analyses, we used a specifically designed software package for UK Biobank, PHESANT (PHEnome Scan Analysis)[25], available in R and ran an automated pre-processing step before developing ML models (Supplementary Methods online). In the below we describe the analyses using the GBDT-SHAP pipeline, with details for comparisons against LASSO and XGBoost with built-in feature importance methods for feature selection described in Supplementary Methods online.

Identifying potential risk factors.

Potential risk factors were identified by following the six steps, namely, (a) learning curve analysis to determine sufficient amount of data for training (b) developing GBDT models with all available predictors and assessing model performance, (c) calculating variable importance using SHAP values and eliminating predictors based on a threshold, (d) further elimination of highly monotonically correlated predictors, (e) learning curve analysis to determine sufficient amount of data for training with the reduced set of predictors and (f) ensuring that the reduced set of predictors is appropriate from model performance perspective as well as from an epidemiological perspective.

Our GBDT models used in the above steps are binary classifiers, that is, their input are the predictors for each individual and their output is model's confidence for mortality status of that individual. The classes were imbalanced (death rate was around 2.9%) and to address the class imbalance problem, all our ML models were developed with the hyperparameter for scaling positive class weight set to the ratio of negative to positive training samples[26,27]. We initially split the entire data into random training, development, and test sets at the ratio of 60:20:20. The training and the development sets were used as the derivation cohort and the test set as the validation cohort. To avoid overfitting to the training data (which is more common in high-dimensional datasets[28]), we used the development set in all our GBDT models for early stopping of training, and thus effectively tuning the hyper parameter for number of estimators. We assessed model performance using area under the receiver operating characteristics (AUROC), a widely used threshold independent metric in assessing binary classifiers. Confidence intervals of AUROC were calculated using 1000 bootstrap[29] datasets based on the test set. We also report sensitivity and specificity at Youden index optimal cut-off point[30]. We used CatBoost version 0.21 implemented in Python (version 3.5.2, Python Software Foundation) for GBDT model development.

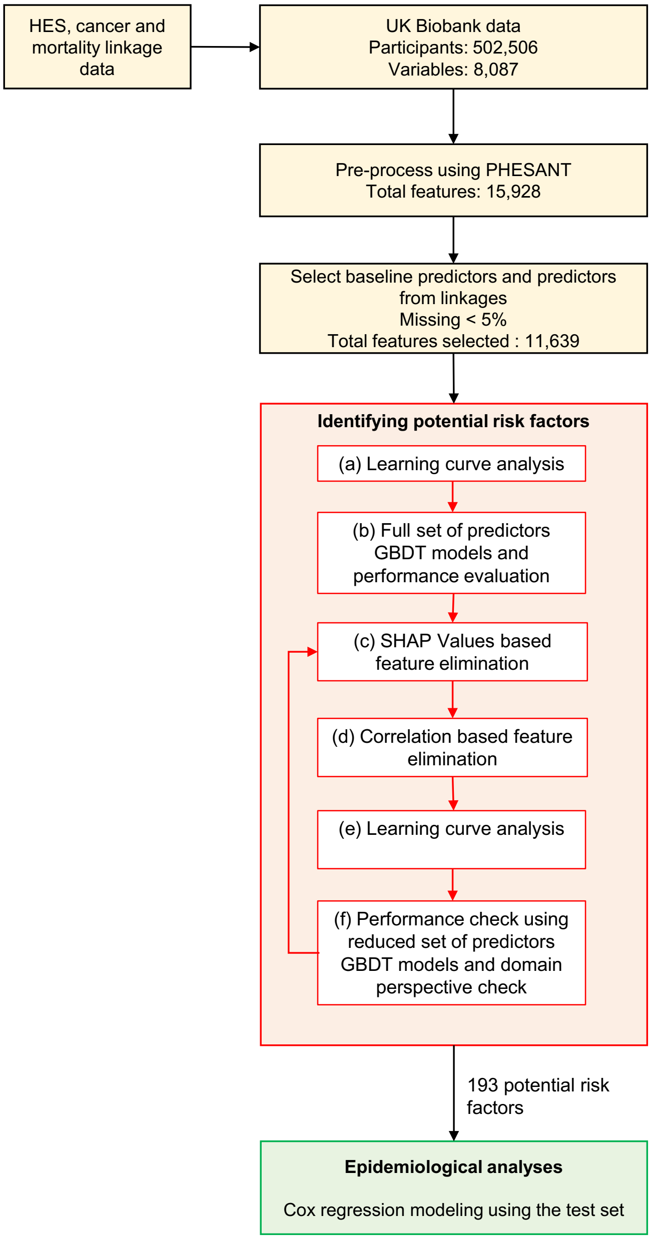

HES, cancer and mortality linkage data
UK Biobank data
Participants: 502,506
Variables: 8,087
Pre-process using PHESANT
Total features: 15,928
Select baseline predictors and predictors from linkages
Missing < 5%
Total features selected : 11,639
Identifying potential risk factors
(a) Learning curve analysis
(b) Full set of predictors GBDT models and performance evaluation
(c) SHAP Values based feature elimination
(d) Correlation based feature elimination
(e) Learning curve analysis
(f) Performance check using reduced set of predictors GBDT models and domain perspective check
193 potential risk factors
Epidemiological analyses
Cox regression modeling using the test set

Figure 1. GBDT-SHAP machine learning pipeline for risk factor discovery, followed by epidemiological analyses using Cox regression. GBDT: gradient boosting decision trees; HES: hospital episode statistics; PHESANT: PHEnome Scan Analysis; SHAP: SHapley Additive exPlanation.

In step (a), we trained GBDT models with increasing numbers of training samples from the training split, starting from 20,000 participants and incremented by 20,000 each time until all training samples were used and verified the adequacy of training samples. Subsequently, we developed GBDT models with all available predictors and assessed their performance in step (b). In step (c), we calculated the importance of each predictor as the mean absolute SHAP value in the training set and normalized variable importance so that they summed up to 100%. We eliminated ‘irrelevant’ predictors using an arbitrarily chosen SHAP threshold. We explored different thresholds to identify ‘important’ predictors and assessed the effects on model performance when using reduced sets of predictors (step (f)). We used Spearman’s ρ (above 0.9) to identify sets of highly correlated predictors and removed all but one (the one recorded for the greatest number of samples) from those sets to produce the final set of predictors for further epidemiological analyses. Such a step avoided carrying forward predictors such as both ‘left leg fat percentage’ and ‘right leg fat percentage’ for further analyses. We repeated the learning curve analysis (step (e)) and model performance check (step (f)) with the reduced set of predictors GBDT models before proceeding to further analyses.

Epidemiological analyses. As the next stage in our proof of principle analyses, we fitted Cox regression models to allow for confounder adjustments and direct interpretation of the associations between the predictors and mortality risk[31]. We fitted univariate Cox models, and also adjusted for baseline predictors identified as 'important' by GBDT, including age, sex, Townsend deprivation index, assessment center and month of birth using the test dataset. We used FDR to account for multiple testing. We present data as Hazard ratios (HR) and their 95% confidence intervals. For selected known mortality predictors which were picked up by GBDT, but not supported by simple Cox modelling, we fitted non-linear models and accounted for selected interactions. We constructed a loop iterating through two-way interactions between predictors which had shown evidence for an association in the GBDT-SHAP pipeline but not confirmed in the Cox models and other predictors identified by GBDT-SHAP pipeline, to further examine mortality associations. All the interaction analyses in the loop were adjusted for baseline predictors as for the other analyses. As our intention was to test for proof of principle in risk factor discovery rather than predictive modelling, for simplicity, we interpret coefficients from Cox models as 'average associations', avoiding the requirement to test for proportionality of hazards assumption. All epidemiological models were done using STATA (version 15, StataCorp, College Station, TX, USA).

Results

Participants characteristics. Of the 502,506 participants included in our study, 14,421 participants (2.9%) died over the median of 7 years (IQR 6.3–7.7 years) of follow-up. Table 1 shows the distribution of the study population and mortality according to selected baseline characteristics. The training and development sets had 8552 and 2941 deaths respectively, while the test set contained 2928 deaths. Participants who died during the follow-up were older and more commonly male compared to those who stayed alive. Those who died during the follow-up period were less educated, had poorer self-rated health, were current or previous smokers and from more deprived backgrounds.

Pre-processing. PHESANT pre-processing, after satisfying our missing value criterion for baseline predictors, derived 11,639 predictors falling under ten broad categories, baseline characteristics, demographics, lifestyle and environment, physical measurements, cognitive function, psychosocial factors, self-reported diseases, medications and operations, health and medical history and hospital diagnoses (Supplementary Table S2 online). Hospital diagnoses (through record linkage) accounted for 98% of the predictors.

Identifying potential risk factors. Our learning curve analysis using all the predictors showed improvements in AUROC as more and more training samples were used (Supplementary Fig. S1 online). We found that in the range of 40% to 60% of samples used for training, model performance stabilized. The GBDT model with all predictors reported an AUROC value of 0.94 (95% CI 0.94–0.95) on the test set (Supplementary Fig. S2 online). The model reported a sensitivity of 0.83 and a specificity of 0.92. At an arbitrary cut-off value of 0.05%, 218 predictors were considered to be 'important'. Correlation based predictor elimination resulted in further reduction of 25 predictors resulting in 193 'important' predictors. Learning curve analysis showed data could be split at the ratio of 60:20:20 also for the reduced set of predictors GBDT models. Reduced set of predictors GBDT model reported an AUROC value of 0.94 (95% CI 0.93–0.95). The model had the same sensitivity and specificity as that of all predictors model.

Figure 2 shows the category-wise predictor importance distribution and Supplementary Table S3 online lists all important predictors. Hospital diagnoses, health and medical history, and self-reported health jointly covered about 60% of the total variable importance summed up, with baseline characteristics (e.g., age, sex), sociodemographics (e.g., employment, education, housing, ethnicity), and lifestyle factors (e.g., smoking, physical activity, diet) each contributing about 10%. Since the mean absolute SHAP values do not directly indicate the direction of association, we show the SHAP summary plot for all important predictors in Supplementary Fig. S3 online.

Epidemiological analyses. In Cox models adjusted for age, sex, Townsend deprivation index, assessment center and month of birth, 166 out of 193 predictors had an association with mortality at $P < 0.05$ after correcting for FDR (Supplementary Table S4 online). Hazard ratios from Cox models of top 50 predictors ranked by SHAP values are shown in Figs. 3 and 4. Supplementary Fig. S4 online shows hazard ratios for all the important predictors. Confirmed predictors included expected mortality associations for various disease outcomes, sociodemographic characteristics, and some lifestyle indicators. After FDR correction, there were 19 predictors which showed evidence of association in the unadjusted models but not in the adjusted models, such as length of time at current address, sensitivity/hurt feelings, worrier/anxious feelings, guilty feelings, risk taking, hearing difficulties, whole body fat-free mass, experiencing of headache and knee pain in last month, diagnoses of inguinal hernia, polyp of colon and gonarthrosis. Eight predictors did not meet the P-value threshold in either unadjusted or adjusted models, including month of birth, comparative height size at the age of ten, cheese intake, handedness, irritability, using a gas fire in winter time, gastro-esophageal reflux disease without esophagitis, and other and unspecified malignant neoplasm of skin of other and unspecified parts of face. As an attempt to understand why these eight predictors may have been picked up by GBDT-SHAP modelling, we next looked in more detail at their associations with mortality. For example, GBDT picked up month of birth as an important predictor and when we recoded it to seasons, both univariate and multivariate models showed modest evidence for an association ($P \leq 0.02$). Interaction loop analyses suggested that association between comparative height at age 10 and mortality might have arisen from an interaction with secondary malignant neoplasm of brain and cerebral meninges ($P_{\text{interaction}} = 1.48\text{E}{-}05$). Similarly, gastro-esophageal reflux disease without esophagitis showed some evidence for interaction with hypertension ($P_{\text{interaction}} = 0.02$) and malignant neoplasm of skin of other and unspecified parts of face had an interaction with fed-up feelings ($P_{\text{interaction}} = 0.009$). Other factors such as cheese

	Total, N (%)	Died, Cases (%)	
Characteristics	**(n = 502,506)**	**(n = 14,421)**	***P*-value***
Age			1.0E−300
< 50	117,874 (23.5)	1056 (0.9)	
50–59.9	167,142 (33.3)	3442 (2.1)	
60–69.9	215,065 (42.8)	9739 (4.5)	
70 +	2425 (0.5)	184 (7.6)	
Sex			4.55E−247
Female	273,384 (54.4)	5668 (2.1)	
Male	229,122 (45.6)	8753 (3.8)	
Ethnic background			2.41E−09
White European	472,697 (94.1)	13,862 (2.9)	
South Asian	9882 (2.0)	170 (1.7)	
East Asian	1574 (0.3)	22 (1.4)	
Black African	8061 (1.6)	115 (1.4)	
Other/mixed	7516 (1.5)	147 (2.0)	
Unknown	2776 (0.6)	105 (3.8)	
Country of birth			1.84E−07
England	390,499 (78.0)	11,059 (2.8)	
Wales	22,072 (4.4)	710 (3.2)	
Scotland	40,176 (8.0)	1536 (3.8)	
Northern Ireland/Republic of Ireland	8068 (1.6)	281 (3.5)	
Elsewhere	39,909 (8.0)	786 (2.0)	
(Missing)	1782 (0.4)	49 (2.8)	
BMI			7.64E−72
Underweight	2626 (0.5)	167 (6.4)	
Normal	162,523 (32.3)	3946 (2.4)	
Overweight	212,065 (42.2)	5742 (2.7)	
Obese	122,187 (24.3)	4313 (3.5)	
(Missing)	3105 (0.6)	253 (8.2)	
Smoking			< 1.00E−300
Non-smokers	273,522 (54.4)	5328 (2.0)	
Ex-smokers	173,058 (34.4)	6042 (3.5)	
Smokers—no type	13,826 (2.8)	458 (3.3)	
Cigars/pipes	2668 (0.5)	208 (7.8)	
Cigarettes < 1–15	22,010 (4.4)	1074 (4.9)	
Cigarettes > 15	14,474 (2.9)	1178 (8.1)	
(Missing)	2948 (0.6)	133 (4.5)	
Qualification			2.20E−59
None	86,037 (17.1)	4,421 (5.1)	
NVQ/CSE/A-levels	175,063 (34.8)	4512 (2.6)	
Degree/professional	235,014 (46.8)	5230 (2.2)	
(Missing)	6392 (1.3)	258 (4.0)	
Townsend deprivation index			0.017
Q1 lowest	125,422 (25.0)	3,016 (2.4)	
Q2	125,516 (25.0)	3182 (2.5)	
Q3	125,468 (25.0)	3527 (2.8)	
Q4 highest	125,477 (25.0)	4682 (3.7)	
(Missing)	623 (0.1)	14 (2.3)	
Overall health rating			< 1.00E−300
Excellent	81,859 (16.3)	1266 (1.6)	
Good	289,016 (57.5)	6318 (2.2)	
Fair	105,367 (21.0)	4362 (4.1)	
Poor	22,777 (4.5)	2280 (10.0)	
(Missing)	3487 (0.7)	195 (5.6)	

Table 1. Baseline characteristics of the UK Biobank cohort. **P*-values are from logistic regression models adjusted for baseline predictors including age, sex, Townsend deprivation index, assessment center and month of birth.

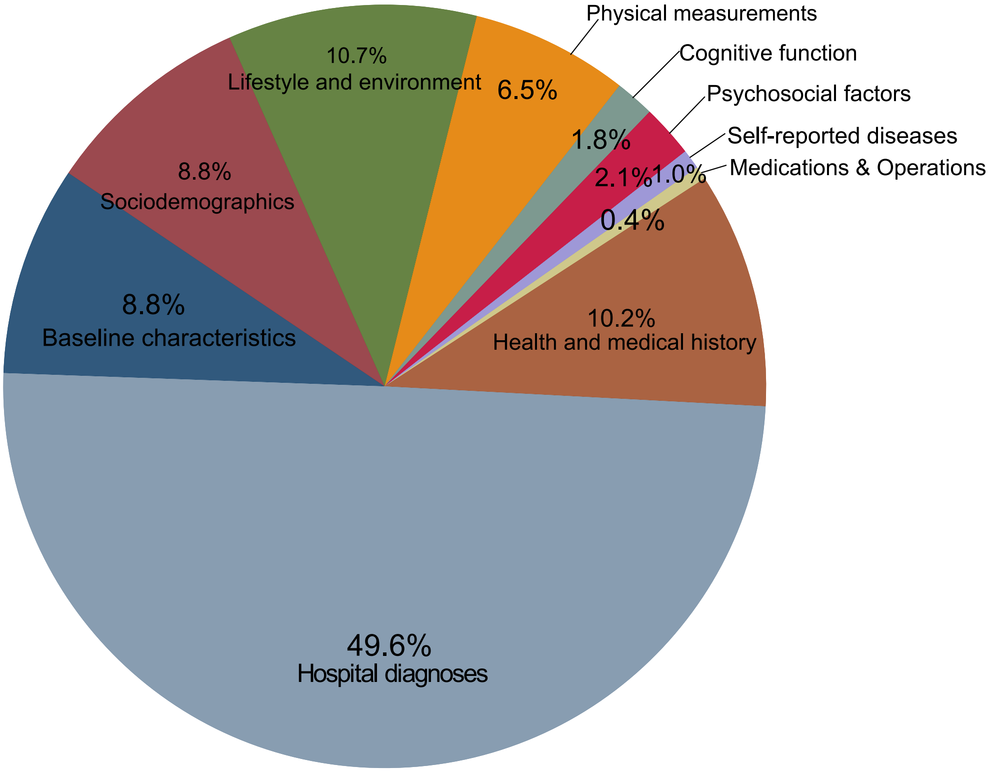


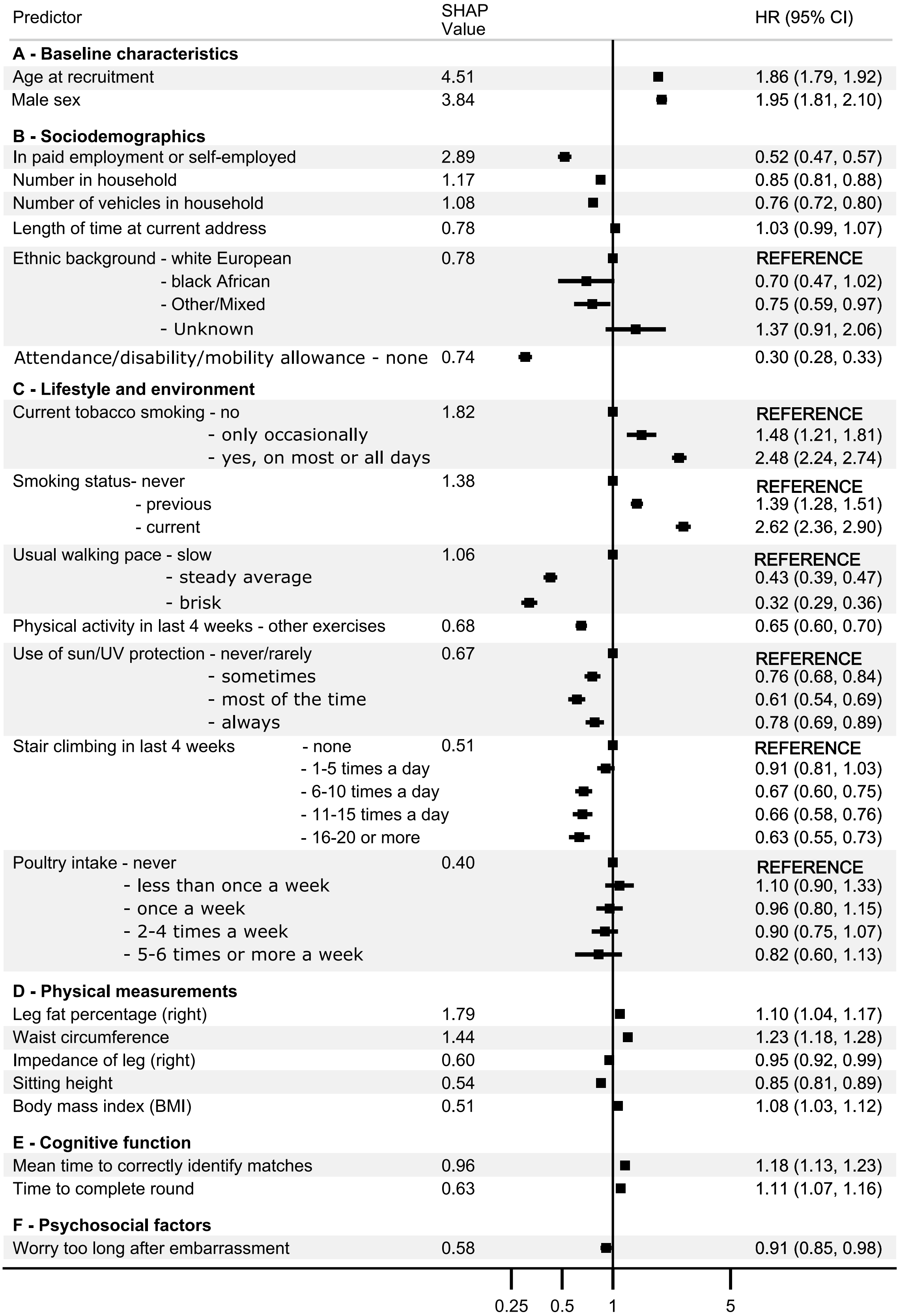
Figure 2. Variable importance values of the 193 important predictors identified for the SHAP value threshold of 0.05%, aggregated into ten categories. Variable importance is calculated as mean absolute SHAP value for each predictor and normalized to 100% before applying the threshold. SHAP: SHapley Additive exPlanation.

intake, handedness, irritability, and gas/solid fuel cooking all had low SHAP values (all < 0.08%). Although univariate Cox model showed an association between tea intake and mortality, an adjusted Cox model showed no evidence for a linear association, ($P = 0.31$). However, there was significant non-linearity ($P_{curvature} = 2.27E{-}10$), with lower mortality for participants drinking 1 to 7 cups per day compared to non-drinkers and the very high intake group (Supplementary Fig. S5 online).

Comparison of feature selection between GBDT-SHAP against LASSO and XGBoost. LASSO performed well in identifying disease associated features (Supplementary Table S5). However, LASSO did not return an association between BMI (or other adiposity indicators), possibly due to a non-linear association with mortality. LASSO did not also select other important predictors such as Townsend deprivation index, or age at cancer diagnosis which were picked up in GBDT-SHAP pipeline. Age at cancer diagnosis is a variable from cancer registration linkage, where information is only included for those participants who had had cancer (84% of the values were missing). While LASSO did not detect the relevance of this information for mortality prediction in the presence of missing information, GBDT-SHAP pipeline was able to rank this feature as the second most important feature. XGBoost was expected to return similar features compared to GBDT-SHAP. However, in analyses of adjusted stability scores, GBDT-SHAP pipeline had higher adjusted features stability scores and highest stability scores, from top 200 features onwards (Supplementary Fig. S7). More importantly, we observed that the default feature importance method of XGBoost, 'weight' (based on number of times a feature was used for splitting in creating decision trees), had poorer adjusted stability scores as compared to GBDT-SHAP (for example, for 200 features selected, 0.86 versus 0.74 and for 250 features selected, 0.88 versus 0.73). LASSO had consistently lower scores compared to GBDT-SHAP values regardless of number of features selected. LASSO's scores (for example, for 200 and 250 features selected, 0.75), were similar to that of XGBoost with default feature importance method.

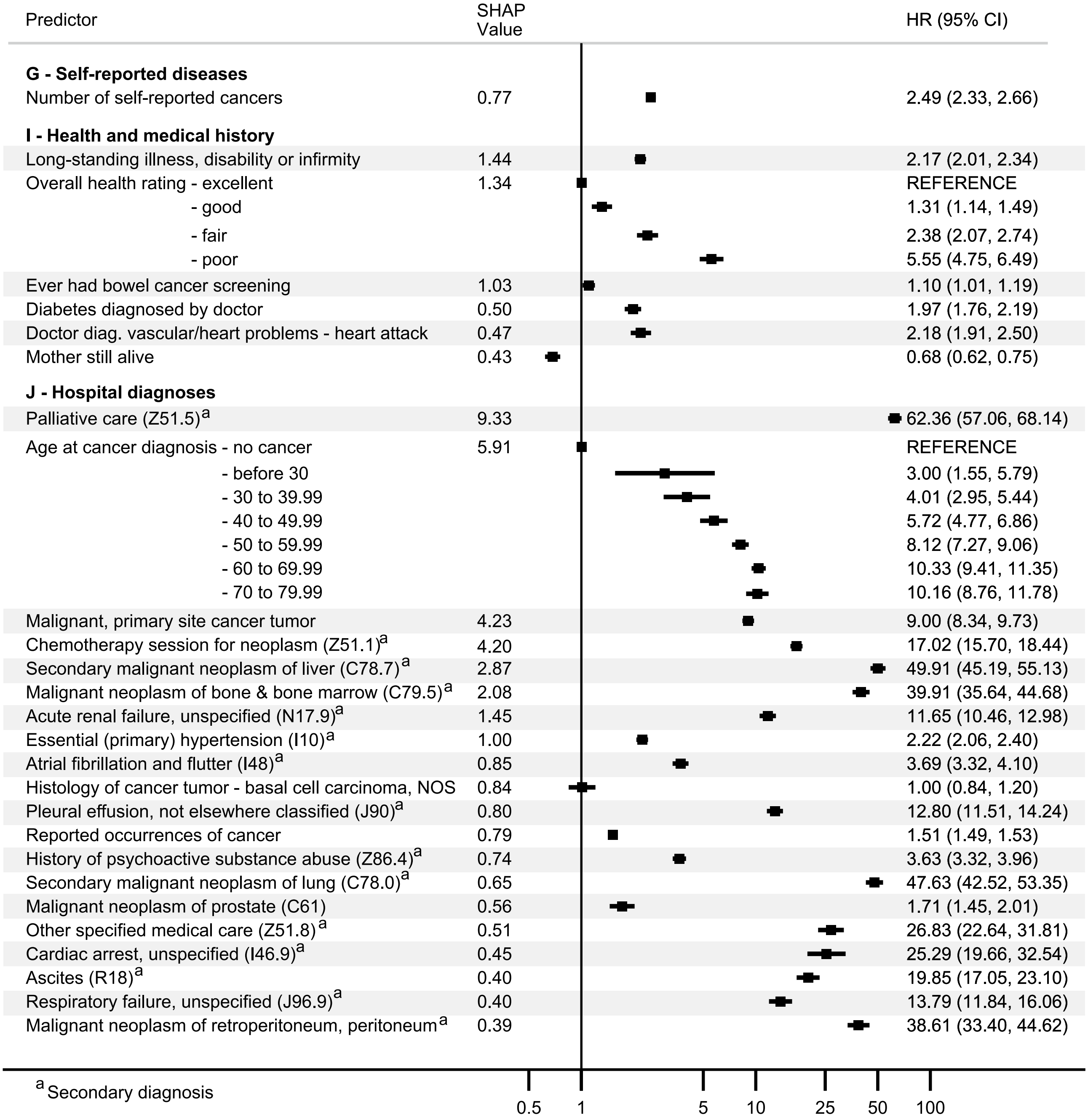
Discussion

We examined the value of GBDT-SHAP pipeline in risk factor discovery using mortality prediction in the UK Biobank as the test case. Our test case picked up the expected predictors (e.g., age, sex, palliative care, disease diagnoses) and many other well-known associations (e.g., smoking, BMI, social differentials). This demonstrates the effectiveness and viability of GBDT-SHAP pipeline for large-scale hypothesis-free screening in this type of multivariable context where standard epidemiological approaches are not feasible. It also provides a better alternative to other approaches such as LASSO, in terms of capturing non-linear predictors or embedded feature selection method such as XGBoost with built-in feature importance, in terms of better feature stability. Feature



Predictor	SHAP Value	HR (95% CI)
A - Baseline characteristics		
Age at recruitment	4.51	1.86 (1.79, 1.92)
Male sex	3.84	1.95 (1.81, 2.10)
B - Sociodemographics		
In paid employment or self-employed	2.89	0.52 (0.47, 0.57)
Number in household	1.17	0.85 (0.81, 0.88)
Number of vehicles in household	1.08	0.76 (0.72, 0.80)
Length of time at current address	0.78	1.03 (0.99, 1.07)
Ethnic background - white European	0.78	REFERENCE
- black African		0.70 (0.47, 1.02)
- Other/Mixed		0.75 (0.59, 0.97)
- Unknown		1.37 (0.91, 2.06)
Attendance/disability/mobility allowance - none	0.74	0.30 (0.28, 0.33)
C - Lifestyle and environment		
Current tobacco smoking - no	1.82	REFERENCE
- only occasionally		1.48 (1.21, 1.81)
- yes, on most or all days		2.48 (2.24, 2.74)
Smoking status- never	1.38	REFERENCE
- previous		1.39 (1.28, 1.51)
- current		2.62 (2.36, 2.90)
Usual walking pace - slow	1.06	REFERENCE
- steady average		0.43 (0.39, 0.47)
- brisk		0.32 (0.29, 0.36)
Physical activity in last 4 weeks - other exercises	0.68	0.65 (0.60, 0.70)
Use of sun/UV protection - never/rarely	0.67	REFERENCE
- sometimes		0.76 (0.68, 0.84)
- most of the time		0.61 (0.54, 0.69)
- always		0.78 (0.69, 0.89)
Stair climbing in last 4 weeks - none	0.51	REFERENCE
- 1-5 times a day		0.91 (0.81, 1.03)
- 6-10 times a day		0.67 (0.60, 0.75)
- 11-15 times a day		0.66 (0.58, 0.76)
- 16-20 or more		0.63 (0.55, 0.73)
Poultry intake - never	0.40	REFERENCE
- less than once a week		1.10 (0.90, 1.33)
- once a week		0.96 (0.80, 1.15)
- 2-4 times a week		0.90 (0.75, 1.07)
- 5-6 times or more a week		0.82 (0.60, 1.13)
D - Physical measurements		
Leg fat percentage (right)	1.79	1.10 (1.04, 1.17)
Waist circumference	1.44	1.23 (1.18, 1.28)
Impedance of leg (right)	0.60	0.95 (0.92, 0.99)
Sitting height	0.54	0.85 (0.81, 0.89)
Body mass index (BMI)	0.51	1.08 (1.03, 1.12)
E - Cognitive function		
Mean time to correctly identify matches	0.96	1.18 (1.13, 1.23)
Time to complete round	0.63	1.11 (1.07, 1.16)
F - Psychosocial factors		
Worry too long after embarrassment	0.58	0.91 (0.85, 0.98)

Figure 3. Adjusted Cox regression hazard ratios (HR) with 95% confidence intervals and SHAP values (normalized for 100%) for top 50 predictors ranked by SHAP values belonging to the categories of baseline characteristics, sociodemographics, lifestyle and environment, physical measurements, cognitive functions, and psychosocial factors. Estimates are adjusted for age, sex, Townsend deprivation index, assessment center, and month of birth. The ethnic group "east Asian" is not shown as it had a hazard ratio of 1.4E−20. SHAP: SHapley Additive exPlanation.



Predictor	SHAP Value	HR (95% CI)
G - Self-reported diseases		
Number of self-reported cancers	0.77	2.49 (2.33, 2.66)
I - Health and medical history		
Long-standing illness, disability or infirmity	1.44	2.17 (2.01, 2.34)
Overall health rating - excellent	1.34	REFERENCE
- good		1.31 (1.14, 1.49)
- fair		2.38 (2.07, 2.74)
- poor		5.55 (4.75, 6.49)
Ever had bowel cancer screening	1.03	1.10 (1.01, 1.19)
Diabetes diagnosed by doctor	0.50	1.97 (1.76, 2.19)
Doctor diag. vascular/heart problems - heart attack	0.47	2.18 (1.91, 2.50)
Mother still alive	0.43	0.68 (0.62, 0.75)
J - Hospital diagnoses		
Palliative care (Z51.5)[a]	9.33	62.36 (57.06, 68.14)
Age at cancer diagnosis - no cancer	5.91	REFERENCE
- before 30		3.00 (1.55, 5.79)
- 30 to 39.99		4.01 (2.95, 5.44)
- 40 to 49.99		5.72 (4.77, 6.86)
- 50 to 59.99		8.12 (7.27, 9.06)
- 60 to 69.99		10.33 (9.41, 11.35)
- 70 to 79.99		10.16 (8.76, 11.78)
Malignant, primary site cancer tumor	4.23	9.00 (8.34, 9.73)
Chemotherapy session for neoplasm (Z51.1)[a]	4.20	17.02 (15.70, 18.44)
Secondary malignant neoplasm of liver (C78.7)[a]	2.87	49.91 (45.19, 55.13)
Malignant neoplasm of bone & bone marrow (C79.5)[a]	2.08	39.91 (35.64, 44.68)
Acute renal failure, unspecified (N17.9)[a]	1.45	11.65 (10.46, 12.98)
Essential (primary) hypertension (I10)[a]	1.00	2.22 (2.06, 2.40)
Atrial fibrillation and flutter (I48)[a]	0.85	3.69 (3.32, 4.10)
Histology of cancer tumor - basal cell carcinoma, NOS	0.84	1.00 (0.84, 1.20)
Pleural effusion, not elsewhere classified (J90)[a]	0.80	12.80 (11.51, 14.24)
Reported occurrences of cancer	0.79	1.51 (1.49, 1.53)
History of psychoactive substance abuse (Z86.4)[a]	0.74	3.63 (3.32, 3.96)
Secondary malignant neoplasm of lung (C78.0)[a]	0.65	47.63 (42.52, 53.35)
Malignant neoplasm of prostate (C61)	0.56	1.71 (1.45, 2.01)
Other specified medical care (Z51.8)[a]	0.51	26.83 (22.64, 31.81)
Cardiac arrest, unspecified (I46.9)[a]	0.45	25.29 (19.66, 32.54)
Ascites (R18)[a]	0.40	19.85 (17.05, 23.10)
Respiratory failure, unspecified (J96.9)[a]	0.40	13.79 (11.84, 16.06)
Malignant neoplasm of retroperitoneum, peritoneum[a]	0.39	38.61 (33.40, 44.62)

[a] Secondary diagnosis

Figure 4. Adjusted Cox regression hazard ratios (HR) with 95% confidence intervals and SHAP values (normalized for 100%) for top 50 predictors ranked by SHAP values belonging to the categories of self-reported diseases, health and medical history and hospital diagnoses. Estimates are adjusted for age, sex, Townsend deprivation index, assessment center, and month of birth. SHAP: SHapley Additive exPlanation.

stability is an important aspect of feature selection for domain experts as it provides assurance to them that the selected features are robust to the perturbation of input data[32]. Also, other studies have shown the inconsistent feature attribution behavior of feature selection method such as gain and split count used in XGBoost[16]. Our approach has particular interest in the context of relatively rare diseases for which we know little about, and where large-scale data now provides the first opportunities to identify candidates for prevention.

ML methods are increasingly used in disease prognosis and there is one previous study using ML to predict all-cause mortality in the UK Biobank. This earlier study compared artificial neural network and random forest methods against Cox regression using a set of 60 variables, selected based on their biological plausibility[7]. As an additional validation, where included in our dataset, our hypothesis-free approach, including over 11,000 predictors in the GBDT-SHAP pipeline, picked up all these risk factors or their equivalent. These included key characteristics, such as age, sex, ethnicity, education, Townsend deprivation index, prior cancer diagnoses, smoking, physical activity, blood pressure, diabetes, and adiposity, confirming that our approach is able to identify relevant indicators ‘hidden’ amongst thousands of predictors. While most of the mortality predictors identified in our models were very logical and expected, in a context when less is known about potential predictors, our comprehensive hypothesis-free approach shows great promise for the identification of novel risk factor candidates.

Although GBDT and other ML models tend to be complex and less interpretable than traditional approaches[33,34], a strength with this approach is the ability to identify relevant risk factors in the context of interactions and non-linear associations. Here, epidemiological analyses using Cox, or any other generalized linear models require careful model construction which is often impractical when dealing with a very large number of predictors, and complex unknown interactions. Data pre-processing requirements are less for GBDT than that required for standard epidemiological approaches, and our analyses provided examples where non-linear associations which would have remained hidden in standard epidemiological analyses were picked up by the GBDT-SHAP pipeline. Another strength is the ability to incorporate information from thousands of predictors, and to better cope with missing information (without having the need to impute using linear approaches such as MICE[35] or non-linear approaches such as MissForest[36]) in this type of multivariable context. However, in this type of real-life setting, GBDT-SHAP approach is unsuitable for simultaneous inclusion of data from the UK Biobank sub-samples collected after the baseline as participation in the follow-up surveys is correlated with mortality. GBDT-SHAP pipeline also picked up indicators which were associated with the outcome purely due to confounding, as shown by associations of several of the identified ‘important’ predictors being explained by a standard adjustment for baseline factors. Confounding and multicollinearity can also notably affect the SHAP based importance ranking. For example, while the number of cancer diagnoses came among the most important factors, SHAP ranked some cancer diagnoses as less important than age related predictors such as experiencing of knee pain last month, sitting height, and gonarthrosis. For risk factor discovery this may have relatively little importance if at least one relevant indicator is picked up, however, this highlights the importance of replication and more detailed modelling, with caution required when interpreting apparent associations without clear explanations.

One of the challenges with our approach arises from the need to account for multiple testing and the lack of a pre-specified cut off value to consider indicators as ‘important’ in the context of feature selection. Some authors suggest that 3% of total number of features could be considered as relevant if the number of predictors included exceeds 100[37]. Our choice was based on a SHAP value threshold (0.05), which selected a slightly smaller proportion of features (1.87%). Our chosen threshold resulted from a pragmatic assessment where we hoped to be inclusive enough to allow for possibly relevant and interesting features be taken forward for further analyses, while at the same time limiting numbers to a manageable amount and not losing too much on the predictive ability of the model. After screening under SHAP threshold, we further used FDR correction threshold to account for multiple testing with the aim of reducing Type I error. Another pragmatic approach which we could consider in this context would be to ignore the pre-screening by SHAP threshold and to use Bonferroni correction based on the total number of features when determining the *P*-value threshold for the epidemiological analyses. However, this would be overly conservative and increase the risk of Type II error, as suggested by the inability to identify well-known mortality risk factors such as the BMI and other adiposity indices. Having said that, even this approach would not have led to findings very dissimilar to those reported, as Bonferroni correction based on the total number of features would have led to 133 features for follow on analyses (compared to 193 with our approach).

Our study demonstrates some of the opportunities in ML based risk factor discovery utilizing the recent implementation of GBDT (CatBoost) and Shapley values-based feature importance method (TreeSHAP). There are also limitations, some of which are specific with respect to the dataset. Indeed, reliable analyses from any model require the understanding of the data from which the results are derived. Here, UK Biobank is a cohort of volunteers with higher education and socio-economic status, and lower mortality rates compared to the general population[38]. This type of healthy volunteer bias may affect the external validity of our findings. However, it was reassuring that our data-driven approach identified the traditional risk factors, suggesting the ability to obtain valuable insights in other, less explored settings of risk prediction. Furthermore, as all the analyses in our study were done using a single dataset, we cannot exclude problems with overidentification. Also, as our model used only those predictors available for at least 95% of the participants, we may have left out important determinants which had not been captured in the full cohort. In this observational method exploration, we also cannot establish causal effects, and as we only included adjustments for very basic covariates in our proof of concept test-case, confounding is likely to explain some of the associations.

In conclusion, our data-driven, hypothesis-free approach utilizing specific ML methods was a viable, fast, and pragmatic approach to risk factor discovery in a highly phenotyped high dimensional tabular data. Our approach was able to pick up traditional risk factors from among thousands of possible predictors and showed potential for discovering relevant mortality predictors in the context of interactions, non-linear associations, and missing values. However, to ensure interpretability of the identified predictor—outcome associations, a more detailed modelling utilizing domain expertise and traditional methods is still required.

Received: 1 July 2021; Accepted: 16 November 2021
Published online: 26 November 2021

References

1. Mitchell, T. M. *Machine Learning* (McGraw-Hill, 1997).
2. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).

3. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PLoS ONE* **12**, e0174944 (2017).
4. Hernesniemi, J. A. *et al.* Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome–the MADDEC study. *Ann. Med.* **51**, 156–163 (2019).
5. Blom, M. C., Ashfaq, A., Sant'Anna, A., Anderson, P. D. & Lingman, M. Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: A retrospective, population-based registry study. *BMJ Open* **9**, e028015 (2019).
6. Mohamadlou, H. *et al.* Multicenter validation of a machine-learning algorithm for 48-h all-cause mortality prediction. *Health Inform. J.* **26**, 1912–1925 (2019).
7. Weng, S. F., Vaz, L., Qureshi, N. & Kai, J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS ONE* **14**, e0214365 (2019).
8. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
9. Cristianini, N. & Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge University Press, 2000).
10. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
11. Bose, N. K. & Liang, P. *Neural Network Fundamentals with Graphs, Algorithms, and Applications* (McGraw-Hill, Inc, 1996).
12. Kim, S. Y. *et al.* A deep learning model for real-time mortality prediction in critically ill children. *Crit. Care* **23**, 279 (2019).
13. Meyer, A. *et al.* Machine learning for real-time prediction of complications in critical care: A retrospective study. *Lancet Respir. Med.* **6**, 905–914 (2018).
14. Olson, R. S., Cava, W., Mustahsan, Z., Varik, A. & Moore, J. H. Data-driven advice for applying machine learning to bioinformatics problems. *Pac. Symp. Biocomput.* **23**, 192–203 (2018).
15. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf.* **31**, 6638–6648 (2018).
16. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. arXiv preprint http://arxiv.org/abs/1802.03888 (2018).
17. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. arXiv preprint http://arxiv.org/abs/1705.07874 (2017).
18. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
19. Hasan, M. M. *et al.* NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning. *Brief. Bioinform.* (2021).
20. Hasan, M. M. *et al.* Meta-i6mA: An interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.* **22**, bbaa202 (2021).
21. Cheng, C.-Y. *et al.* Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. *Nat. Commun.* **12**, 1–15 (2021).
22. Song, X. *et al.* Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J. Am. Med. Inform. Assoc.* **26**, 242–253 (2019).
23. Sudlow, C. *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**, e1001779. https://doi.org/10.1371/journal.pmed.1001779 (2015).
24. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
25. Millard, L. A., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software application profile: PHESANT: A tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **1**, 123 (2017).
26. Ling, C. X. & Sheng, V. S. Cost-sensitive learning and the class imbalance problem. *Encycl. Mach. Learn.* **2011**, 231–235 (2008).
27. Japkowicz, N. & Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **6**, 429–449 (2002).
28. Maldonado, S., Weber, R. & Famili, F. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Inf. Sci.* **286**, 228–246 (2014).
29. Efron, B. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **82**, 171–185 (1987).
30. Perkins, N. J. & Schisterman, E. F. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.* **163**, 670–675 (2006).
31. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N. & Fanos, V. Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina* **56**, 455 (2020).
32. Kalousis, A., Prados, J. & Hilario, M. Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowl. Inf. Syst.* **12**, 95–116 (2007).
33. Beam, A. L. & Kohane, I. S. Big data and machine learning in health care. *JAMA* **319**, 1317–1318 (2018).
34. Freitas, A. A. Comprehensible classification models: A position paper. *ACM SIGKDD Explor. Newsl.* **15**, 1–10 (2014).
35. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30**, 377–399 (2011).
36. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
37. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. Feature selection for high-dimensional data. *Progr. Artif. Intell.* **5**, 65–75 (2016).
38. Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

Acknowledgements

The authors thank the participants of UK Biobank who answered surveys and made this work possible. I. Madakkatel acknowledges the support of an Australian Government Research Training Program (RPT) Scholarship.

Author contributions

I.M. managed data, conducted ML modelling and statistical analyses, and wrote the first draft. A.Z. advised on statistical analyses and the data. M.M. supervised the study, designed ML experiments, and revised the paper. E.H. conceptualized, supervised, and funded the study, advised on statistical analyses, and revised the paper. All authors interpreted results and approve the final version for submission.

Funding

This project was supported by the National Health and Medical Research Foundation, Australia (GNT1123603).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-02476-9.

Correspondence and requests for materials should be addressed to I.M. or E.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.