ELSEVIER

Journal of Clinical Epidemiology 122 (2020) 56−69

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Logistic regression was as good as machine learning for predicting major chronic diseases

Simon Nusinovici[a], Yih Chung Tham[a,c], Marco Yu Chak Yan[a], Daniel Shu Wei Ting[a,c], Jialiang Li[a,d], Charumathi Sabanayagam[a,c], Tien Yin Wong[a,b,c], Ching-Yu Cheng[a,b,c,*]

[a]*Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore*
[b]*Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*
[c]*Ophthalmology and Visual Sciences Academic Clinical Programme, Duke-NUS Medical School, Singapore*
[d]*Department of Statistics and Applied Probability, National University of Singapore, Singapore*

Accepted 4 March 2020; Published online 10 March 2020

Abstract

Objective: To evaluate the performance of machine learning (ML) algorithms and to compare them with logistic regression for the prediction of risk of cardiovascular diseases (CVDs), chronic kidney disease (CKD), diabetes (DM), and hypertension (HTN) and in a prospective cohort study using simple clinical predictors.

Study Design and Setting: We conducted analyses in a population-based cohort study in Asian adults (n = 6,762). Five different ML models were considered—single-hidden-layer neural network, support vector machine, random forest, gradient boosting machine, and k-nearest neighbor—and were compared with standard logistic regression.

Results: The incidences at 6 years of CVD, CKD, DM, and HTN cases were 4.0%, 7.0%, 9.2%, and 34.6%, respectively. Logistic regression reached the highest area under the receiver operating characteristic curve for CKD (0.905 [0.88, 0.93]) and DM (0.768 [0.73, 0.81]) predictions. For CVD and HTN, the best models were neural network (0.753 [0.70, 0.81]) and support vector machine (0.780 [0.747, 0.812]), respectively. However, the differences with logistic regression were small (less than 1%) and nonsignificant. Logistic regression, gradient boosting machine, and neural network were systematically ranked among the best models.

Conclusion: Logistic regression yields as good performance as ML models to predict the risk of major chronic diseases with low incidence and simple clinical predictors.

Keywords: Machine learning; Logistic regression; Prognostic modeling; Chronic diseases; Interaction; Nonlinearity

1. Introduction

Artificial intelligence using machine learning (ML) is an ensemble of techniques that automatically learn patterns from data and that require no assumptions regarding the structure of the data. ML methods include notably neural networks, support vector machine, random forest, and gradient boosting machine. A strength of these techniques is that they capture nonlinear relationships in the data, as well as interaction between predictors. Many studies have demonstrated their promising performance for diseases prediction [1−7].

However, some studies have shown that the incremental predictive performance beyond standard methods (such as regression) might be limited [8−11]. Others even show no benefit of ML over classical statistical models such as logistic regression [12−17]. This could be due to the amount of data required to reach adequate performance prediction for ML. A simulation study has indeed shown that ML techniques are data hungry and might therefore not be appropriate for small data sets or data sets with a small number of events of interest [18]. Some studies only present the predictive performance of ML models without testing simpler methods [19−21].

Consequently, it is not clear whether ML should be used for disease risk modeling, where the number of incident

Conflict of interest statement: The authors confirm that they have no conflicts of interest associated with this publication.

Funding: This work was supported by the National Medical Research Council [DYNAMO - NMRC/OFLCG/001/2017]. The funding source had no role in study design; in the collection, analysis, or interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

* Corresponding author: Singapore Eye Research Institute, 20 College Road, The Academia, Level 6, Singapore 169856, Singapore. Tel.: +65 6576 7277; fax: +65 6225 2568.

E-mail address: chingyu.cheng@duke-nus.edu.sg (C.-Y. Cheng).

https://doi.org/10.1016/j.jclinepi.2020.03.002
0895-4356/

What is new?

Key findings:

- Logistic regression yields as good performance as machine learning (ML) models to predict the risk of major chronic diseases with low incidence and simple clinical predictors in a prospective epidemiological study of moderate sample size (~10,000 participants at baseline), typical of many epidemiological studies.

- Logistic regression, gradient boosting machine, and neural network were systematically ranked among the best models.

What this adds to what is known?

- It is not clear whether ML is superior to conventional regression for disease prediction modeling, where the number of incident disease cases is low.

What is the implication, what should change now?

- We suggest that traditional regression models should continue to have a key role in disease risk prediction when using a limited number of simple clinical predictors, and the use of ML techniques may not be warranted in such studies

cases is generally quite low. Indeed, many chronic diseases of importance in terms of public health and overall burden have low incidences, such end-stage renal disease [22] or cardiovascular diseases [23]. There is a real need to understand the performance of ML for prediction for these diseases with low incidence (or low number of events).

The objective of this study was to evaluate the performance of ML algorithms and to compare them with the traditional regression technique for the prediction of risk of a range of common chronic diseases including cardiovascular diseases (CVDs), chronic kidney disease (CKD), diabetes (DM), and hypertension (HTN) in a prospective cohort study of moderate sample size, typical of many epidemiological studies. Our hypothesis is that ML methods in case of moderate sample size with a limited number of incident events and simple clinical predictors do not outperform traditional methods such as logistic regression.

2. Materials and methods

2.1. Study population

We conducted analyses in the Singapore Epidemiology of Eye Disease, a population-based prospective cohort study of eye diseases in Asian Chinese, Indian, and Malay adults aged 40-80 years. Details of the study participants and methods have been reported elsewhere [24–27]. Briefly, participants of the three ethnic groups at baseline were randomly selected using an age-stratified sampling method. The baseline study was conducted between 2004 and 2011 on a total of 10,033 participants. The follow-up examinations were conducted ~6 years (6.13 ± 0.96 years) after the baseline visit between 2011 and 2017 on 6,762 participants (response rate 78% among eligible participants). Ineligible participants included people who died, who migrated, and who were prisoners between baseline and follow-up visit. Participants underwent a standardized interview, clinical examination, and laboratory investigations at both baseline and follow-up visits. Informed, written consent was obtained from participants, and ethical approval was obtained from the Institutional Review Board of the Singapore Eye Research Institute.

2.2. Outcome definitions

Four outcomes were considered: CVD, CKD, DM, and HTN. A separate analysis was performed for each outcome. Only new cases (incident cases) reported at the follow-up visit were considered. The prevalent cases at baseline were excluded from the analysis. The definitions considered in this study were the following:

- Incident cases of CVD: self-reported using an interview-based, standardized questionnaire, defined as a history of myocardial infarction, angina pectoris, or stroke. A reliability assessment showed that 75.5% of those who reported a history of CVD at baseline confirmed this information in the follow-up visit [28].
- Incident cases of CKD: estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73 m^2 at the follow-up visit. eGFR was estimated from serum creatinine using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation to account for variations in creatinine levels due to age, gender, and body weight [29].
- Incident cases of DM: nonfasting plasma glucose $\geq$ 11.1 mmol/l or blood glycosylated hemoglobin (HbA1c) $>$ 6.5% or the use of diabetic medications with a previous physician's diagnosis.
- Incident cases of HTN: systolic blood pressure $\geq$ 140 mmHg or diastolic blood pressure $\geq$ 90 mmHg), self-reported history, or use of hypertensive medications.

2.3. Predictors considered

For each outcome, a set of predictors at baseline was considered based on clinical expertise and on literature [30–32].

- CVD: age (continuous), gender (binary), diabetes (binary), HbA1c (continuous), systolic blood pressure (continuous), ethnicity (categorical), income (categorical), educational level (categorical), body mass index

(BMI) (continuous), smoking status (categorical), alcohol consumption (binary), antihypertensive drug (binary), anticholesterol drug (binary), antidiabetes drug (binary), sign of retinopathy (binary), ocular vascular calibers (arteriolar and venular) (both continuous).

- CKD: age (continuous), gender (binary), diabetes (binary), CVD (binary), albumin-to-creatinine ratio (continuous, log$_{10}$ transformed), eGFR (calculated based on serum creatinine using CKD-EPI equation [29], continuous), HbA1c (continuous), systolic blood pressure (continuous), ethnicity (categorical), income (categorical), educational level (categorical), BMI (continuous), smoking status (categorical), alcohol consumption (binary), antihypertensive drug (binary), anticholesterol drug (binary), antidiabetes drug (binary), sign of retinopathy (binary), ocular vascular calibers (arteriolar and venular) (both continuous).
- DM: age (continuous), gender (binary), BMI (continuous), CVD (binary), smoking status (categorical), family history of DM (binary), alcohol consumption (binary), systolic blood pressure (continuous), ethnicity (categorical), income (categorical), educational level (categorical), antihypertensive drug (binary), anticholesterol drug (binary), nonfasting blood glucose (continuous), low-density lipoprotein cholesterol (continuous), high-density lipoprotein cholesterol (continuous), sign of retinopathy (binary), ocular vascular calibers (arteriolar and venular) (both continuous).
- HTN: age (continuous), gender (binary), diabetes (binary), HbA1c (continuous), systolic blood pressure (continuous), diastolic blood pressure (continuous), ethnicity (categorical), income (categorical), educational level (categorical), BMI (continuous), smoking status (categorical), alcohol consumption (binary), anticholesterol drug (binary), antidiabetes drug (binary), hyperlipidemia (binary), sign of retinopathy (binary), ocular vascular calibers (arteriolar and venular) (both continuous).

2.4. Modeling strategy

All the analyses were performed using R software, version 3.5.1 (R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/), and 'caret' R package was used to compare the performance of the models [33]. Furthermore, because of the poor quality of reporting of methodology and findings in several studies [12,34,35], we have used the TRIPOD checklist (Supplementary Material) to ensure a transparent reporting of the methods used [34].

2.4.1. Preprocessing

First, the missing data were imputed using a nonparametric imputation method based on random forest [36] that can cope with nonlinear relations and complex interactions

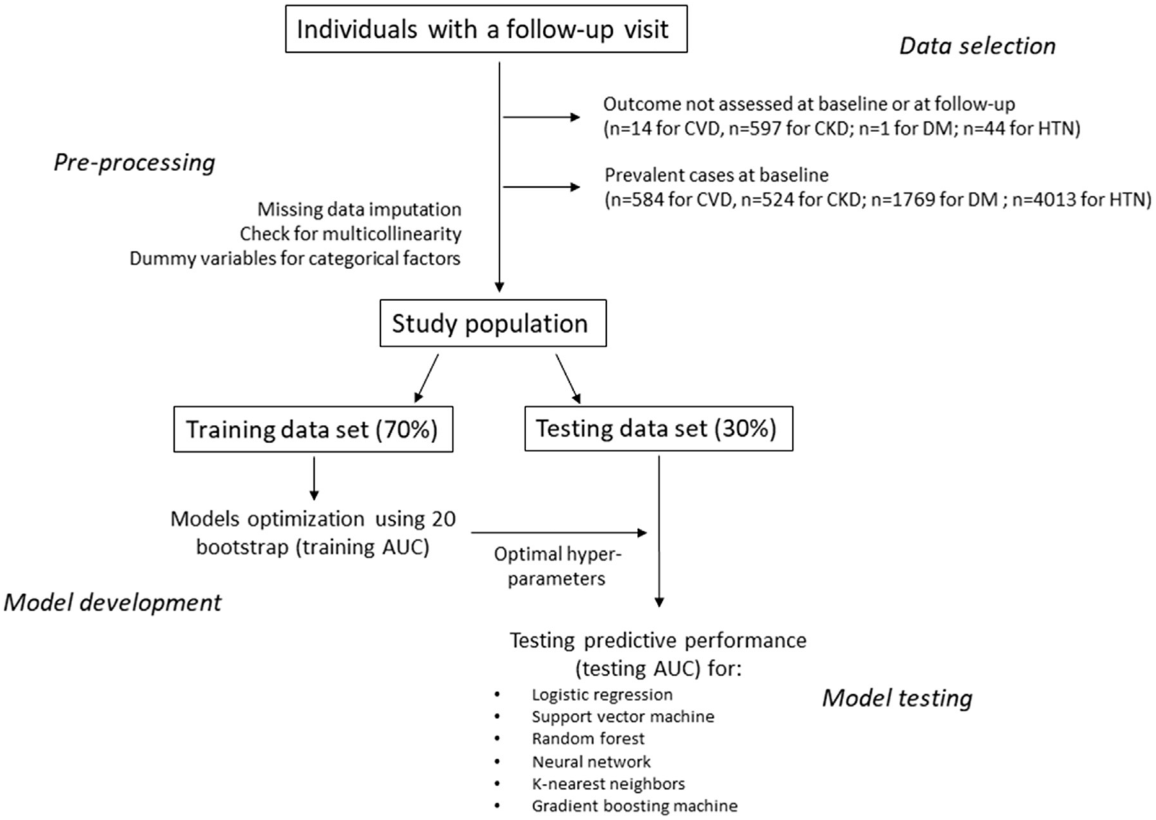
Individuals with a follow-up visit
Data selection
Outcome not assessed at baseline or at follow-up
(n=14 for CVD, n=597 for CKD; n=1 for DM; n=44 for HTN)
Pre-processing
Prevalent cases at baseline
(n=584 for CVD, n=524 for CKD; n=1769 for DM ; n=4013 for HTN)
Missing data imputation
Check for multicollinearity
Dummy variables for categorical factors
Study population
Training data set (70%)
Testing data set (30%)
Models optimization using 20
bootstrap (training AUC)
Optimal hyper-
parameters
Model development
Testing predictive performance
(testing AUC) for:
• Logistic regression
• Support vector machine
• Random forest
• Neural network
• K-nearest neighbors
• Gradient boosting machine
Model testing

Fig. 1. Data selection, preprocessing, model development, and model testing. CVD, cardiovascular diseases; CKD, chronic kidney disease; DM, diabetes; HTN, hypertension.

Table 1. Characteristics of the populations at baseline as per the cardiovascular disease (CVD) status at follow-up (n = 6,164)

	No CVD	CVD	
Characteristics	**5,917**	**247**	***P* Value**
Age			
Median, IQR	55.2 (49.1,63)	60.3 (51.3,66.7)	<0.001
Gender			
Male (*n*, %)	2,703 (45.7)	157 (63.6)	<0.001
Diabetes			
Yes (*n*, %)	1,395 (23.6)	96 (38.9)	<0.001
Glycosylated hemoglobin (%)			
Median, IQR	5.8 (5.6,6.2)	6.1 (5.7,7)	<0.001
Systolic blood pressure (mmHg)			
Median, IQR	135 (122.5,148.5)	144 (130.5,162)	<0.001
Ethnicity			
Chinese (*n*, %)	2,473 (41.8)	39 (15.8)	<0.001
Indians (*n*, %)	1,847 (31.2)	81 (32.8)	
Malay (*n*, %)	1,597 (27)	127 (51.4)	
Income (S$)			
Income <1,000	2,787 (47.1)	144 (58.3)	<0.001
1,000 ≤income <2,000	1,463 (24.7)	61 (24.7)	
2,000 ≤income <3,000	730 (12.3)	27 (10.9)	
Income ≥3,000	937 (15.8)	15 (6.1)	
Educational level			
No formal educational or primary	3,191 (53.9)	177 (71.7)	<0.001
O/N/A levels	2,274 (38.4)	64 (25.9)	
University	452 (7.6)	6 (2.4)	
BMI (kg/m^2)			
BMI (median, IQR)	24.7 (22.2,27.6)	25.8 (23.1,28.8)	<0.001
Smoking status			
Never smoked	4,427 (74.8)	129 (52.2)	<0.001
Current smoker	802 (13.6)	67 (27.1)	
Past smoker	688 (11.6)	51 (20.6)	
Alcohol consumption			
Yes (*n*, %)	517 (8.7)	26 (10.5)	0.331
Antihypertensive drug			
Yes (*n*, %)	1,697 (28.7)	94 (38.1)	0.001
Anticholesterol drug			
Yes (*n*, %)	1,217 (20.6)	55 (22.3)	0.518
Antidiabetes drug			
Yes (*n*, %)	834 (14.1)	59 (23.9)	<0.001
Retinopathy			
Yes (*n*, %)	542 (9.2)	53 (21.5)	<0.001
Arteriolar vessel caliber (μm)			
Median, IQR	134.4 (127.5,142.1)	136 (130.7,143.5)	0.01
Venular vessel caliber (μm)			
Median, IQR	196.4 (187.3,208.1)	201.9 (191.6,212.4)	<0.001

Abbreviations: BMI, body mass index; IQR, interquartile range.

[37]. This method provides highly accurate predictions and has outperformed common techniques such as nearest neighbor imputation and multivariate imputation by chained equations in two clinical predictive models [38]. Second, the categorical predictors (education level, smoking status, income, and ethnicity) were transformed into dummy variables. Third, the continuous predictors were centered (subtract mean from values) and scaled (divide

Table 2. Characteristics of the populations at baseline as per the chronic kidney disease (CKD) at follow-up (n = 5,641)

	No CKD	CKD	
Characteristics	**5,245**	**396**	***P* Value**
Age			
Median, IQR	54 (48.7,61.3)	65.7 (59.7,71.5)	<0.001
Gender			
Male (*n*, %)	2,511 (47.9)	223 (56.3)	0.001
Diabetes			
Yes (*n*, %)	1,154 (22)	211 (53.3)	<0.001
CVD			
Yes (*n*, %)	349 (6.7)	69 (17.4)	<0.001
Albumin-to-creatinine ratio			
Median, IQR	15.2 (7.9,28.7)	34.4 (14.1,91.4)	<0.001
Glomerular filtration rate			
Median, IQR	92.6 (80.7,101.8)	71.2 (65.3,79.9)	<0.001
Glycosylated hemoglobin (%)			
Median, IQR	5.8 (5.6,6.2)	6.4 (5.8,7.4)	<0.001
Systolic blood pressure (mmHg)			
Median, IQR	133 (121.5,146)	149.2 (136.5,164)	<0.001
Ethnicity			
Chinese (*n*, %)	2,105 (40.1)	137 (34.6)	<0.001
Indians (*n*, %)	1,796 (34.2)	110 (27.8)	
Malay (*n*, %)	1,344 (25.6)	149 (37.6)	
Income (S$)			
Income <1,000	2,238 (42.7)	268 (67.7)	<0.001
1,000 ≤income <2,000	1,393 (26.6)	77 (19.4)	
2,000 ≤income <3,000	723 (13.8)	27 (6.8)	
Income ≥3,000	891 (17)	24 (6.1)	
Educational level			
No formal educational or primary	2,675 (51)	281 (71)	<0.001
O/N/A levels	2,139 (40.8)	98 (24.7)	
University	431 (8.2)	17 (4.3)	
BMI (kg/m^2)			
BMI (median, IQR)	24.7 (22.2,27.7)	25.7 (23,28.8)	<0.001
Smoking status			
Never smoked	3,818 (72.8)	265 (66.9)	<0.001
Current smoker	793 (15.1)	57 (14.4)	
Past smoker	634 (12.1)	74 (18.7)	
Alcohol consumption			
Yes (*n*, %)	509 (9.7)	31 (7.8)	0.221
Antihypertensive drug			
Yes (*n*, %)	1,370 (26.1)	235 (59.3)	<0.001
Anticholesterol drug			
Yes (*n*, %)	1,042 (19.9)	159 (40.2)	<0.001
Antidiabetes drug			
Yes (*n*, %)	646 (12.3)	157 (39.6)	<0.001
Retinopathy			
Yes (*n*, %)	433 (8.3)	85 (21.5)	<0.001
Arteriolar vessel caliber (μm)			
Median, IQR	134.3 (127.4,142)	136.6 (128.5,144.1)	0.006
Venular vessel caliber (μm)			
Median, IQR	196.4 (187.3,207.6)	199.8 (189.5,211.3)	0.002

Abbreviations: BMI, body mass index; CVD, cardiovascular disease; IQR, interquartile range.

Table 3. Characteristics of the populations at baseline as per the diabetes (DM) status at follow-up (*n* = 4,992)

	No DM	**DM**	
Characteristics	**4,531**	**461**	***P* Value**
Age			
Median, IQR	54.6 (48.9,62.6)	54.6 (48.5,61.5)	0.097
Gender			
Male (*n*, %)	2,135 (47.1)	231 (50.1)	0.221
CVD			
Yes (*n*, %)	274 (6)	36 (7.8)	0.135
Systolic blood pressure (mmHg)			
Median, IQR	133 (120.5,146.5)	138 (126.5,153.5)	<0.001
Ethnicity			
Chinese (*n*, %)	2,113 (46.6)	121 (26.2)	<0.001
Indians (*n*, %)	1,216 (26.8)	195 (42.3)	
Malay (*n*, %)	1,202 (26.5)	145 (31.5)	
Income (S$)			
Income <1,000	2,060 (45.5)	215 (46.6)	0.521
1,000 ≤income <2,000	1,151 (25.4)	111 (24.1)	
2,000 ≤income <3,000	575 (12.7)	67 (14.5)	
Income ≥3,000	745 (16.4)	68 (14.8)	
Educational level			
No formal educational or primary	2,360 (52.1)	256 (55.5)	0.347
O/N/A levels	1,800 (39.7)	172 (37.3)	
University	371 (8.2)	33 (7.2)	
BMI (kg/m^2)			
BMI (median, IQR)	24.1 (21.8,26.8)	26.6 (23.9,29.6)	<0.001
Smoking status			
Never smoked	3,301 (72.9)	330 (71.6)	0.149
Current smoker	685 (15.1)	62 (13.4)	
Past smoker	545 (12)	69 (15)	
Alcohol consumption			
Yes (*n*, %)	423 (9.3)	43 (9.3)	0.995
Family history of DM			
Yes (*n*, %)	1,678 (37)	241 (52.3)	<0.001
Antihypertensive drug			
Yes (*n*, %)	1,124 (24.8)	143 (31)	0.003
Anticholesterol drug			
Yes (*n*, %)	734 (16.2)	108 (23.4)	<0.001
Retinopathy			
Yes (*n*, %)	209 (4.6)	28 (6.1)	0.16
Arteriolar vessel caliber (μm)			
Median, IQR	133.8 (127.1,141.6)	134.8 (128.1,142.3)	0.111
Venular vessel caliber (μm)			
Median, IQR	195.9 (187.2,207.4)	198.7 (189.6,209.9)	0.002
Nonfasting blood glucose			
Median, IQR	5.2 (4.8,5.9)	5.9 (5.2,6.8)	<0.001
LDL cholesterol			
Median, IQR	3.4 (2.9,4)	3.5 (3,4)	0.154
HDL cholesterol			
Median, IQR	1.2 (1,1.5)	1.1 (0.9,1.3)	<0.001

Abbreviations: BMI, body mass index; CVD, cardiovascular disease; HDL, high-density lipoprotein; IQR, interquartile range; LDL, low-density lipoprotein.

Table 4. Characteristics of the populations at baseline as per the hypertension (HTN) status at follow-up (n = 2,705)

	No HTN	HTN	
Characteristics	**1,769**	**936**	***P* Value**
Age			
Median, IQR	50.2 (46.9,55.2)	54 (48.8,60.6)	<0.001
Gender			
Male (*n*, %)	786 (44.4)	448 (47.9)	0.088
Diabetes			
Yes (*n*, %)	197 (11.1)	197 (21)	<0.001
Glycosylated hemoglobin (%)			
Median, IQR	5.7 (5.5,6)	5.8 (5.6,6.2)	<0.001
Systolic blood pressure (mmHg)			
Median, IQR	119.5 (111.5,127)	129.5 (123,134.5)	<0.001
Diastolic blood pressure (mmHg)			
Median, IQR	71.5 (67,77.5)	76 (70,81)	<0.001
Ethnicity			
Chinese (*n*, %)	756 (42.7)	361 (38.6)	0.084
Indians (*n*, %)	613 (34.7)	337 (36)	
Malay (*n*, %)	400 (22.6)	238 (25.4)	
Income (S$)			
Income <1,000	591 (33.4)	424 (45.3)	<0.001
1,000 ≤income <2,000	496 (28)	228 (24.4)	
2,000 ≤income <3,000	285 (16.1)	127 (13.6)	
Income ≥3,000	397 (22.4)	157 (16.8)	
Educational level			
No formal educational or primary	708 (40)	472 (50.4)	<0.001
O/N/A levels	851 (48.1)	388 (41.5)	
University	210 (11.9)	76 (8.1)	
BMI (kg/m^2)			
BMI (median, IQR)	23.5 (21.2,26.1)	24.6 (22,27.7)	<0.001
Smoking status			
Never smoked	1,298 (73.4)	663 (70.8)	0.089
Current smoker	305 (17.2)	160 (17.1)	
Past smoker	166 (9.4)	113 (12.1)	
Alcohol consumption			
Yes (*n*, %)	150 (8.5)	84 (9)	0.663
Anticholesterol drug			
Yes (*n*, %)	146 (8.3)	119 (12.7)	<0.001
Antidiabetes drug			
Yes (*n*, %)	85 (4.8)	112 (12)	<0.001
Hyperlipidemia			
Yes (*n*, %)	516 (29.2)	324 (34.6)	0.004
Retinopathy			
Yes (*n*, %)	72 (4.1)	75 (8)	<0.001
Arteriolar vessel caliber (μm)			
Median, IQR	136.7 (131.1,144.3)	134.7 (128.3,142.6)	<0.001
Venular vessel caliber (μm)			
Median, IQR	197 (189.2,207.8)	196.4 (188.2,207.5)	0.55

Abbreviations: BMI, body mass index; IQR, interquartile range.

values by standard deviation) for the following models: neural network, support vector machine, and k-nearest neighbor. Finally, the existence of pairwise correlations among the predictors higher than 0.75 was checked. If existing, an algorithm was used to find the minimal set of predictors that can be removed so that the pairwise correlations are below this threshold [33].

2.4.2. Models considered

Six different models were considered: logistic regression, single-hidden-layer neural network (hyperparameters: the number of units in the single layer and decay parameter), radial basis support vector machine (hyperparameter: C-parameter or cost), random forest (hyperparameter: the number of randomly selected variables per node), gradient boosting machine (hyperparameters: shrinkage parameter, the number of trees, and maximum interaction depth), and k-nearest neighbor (hyperparameter: the number of neighbors). The logistic regression can be considered as an ML model. However, we have chosen the dichotomy logistic regression vs. ML because the logistic regression does not require the optimization of any hyperparameter and is thus easier to implement.

2.4.3. Resampling strategy

The following resampling strategy was considered (the same for all the outcomes). First, each data set (corresponding to each outcome) was randomly split into training (70%) and testing (30%) data sets, stratified on the incidence of the cases (to have the same incidence of cases in the training and testing data sets). A bootstrap procedure repeated 20 times was performed in the training data set to find the optimal hyperparameters. The method was chosen because it has a very low variance and is thus appropriate when the goal is to choose between models [39]. Then, the model with the optimal hyperparameter(s) was run in the entire training set and used to predict the risk of cases in the testing data set.

2.4.4. Predictive performance

The predictive performance was considered based on the area under the receiver operating characteristic curve (AUC) with their 95% confidence intervals (CIs), the sensitivity, and the specificity. The thresholds for the cutoff were calculated using Youden's index [40]. The AUC was calculated both in the training data set (bootstrap or training AUC) and in the testing data set (testing AUC). Comparing the training and testing AUC allows us to evaluate the risk of overfitting—higher values of training AUC would suggest the existence of overfitting. For both training and testing AUC, the 95% CI were computed with 2,000 stratified bootstrap replicates. The testing AUC values corresponding to the different models were compared using paired Delong's test [41]. Finally, the variable importance was calculated for each predictor to rank their relative influence on the risk of onset of cases.

2.4.5. Interaction and nonlinearity

The amount of interaction (estimated as the maximum interaction depth) was assessed for each set of predictor using gradient boosting machine. Possible nonlinear relationships between predictors and the outcomes were formally tested by adding squared terms of the continuous predictors in a logistic regression model. Predictors with P-value <

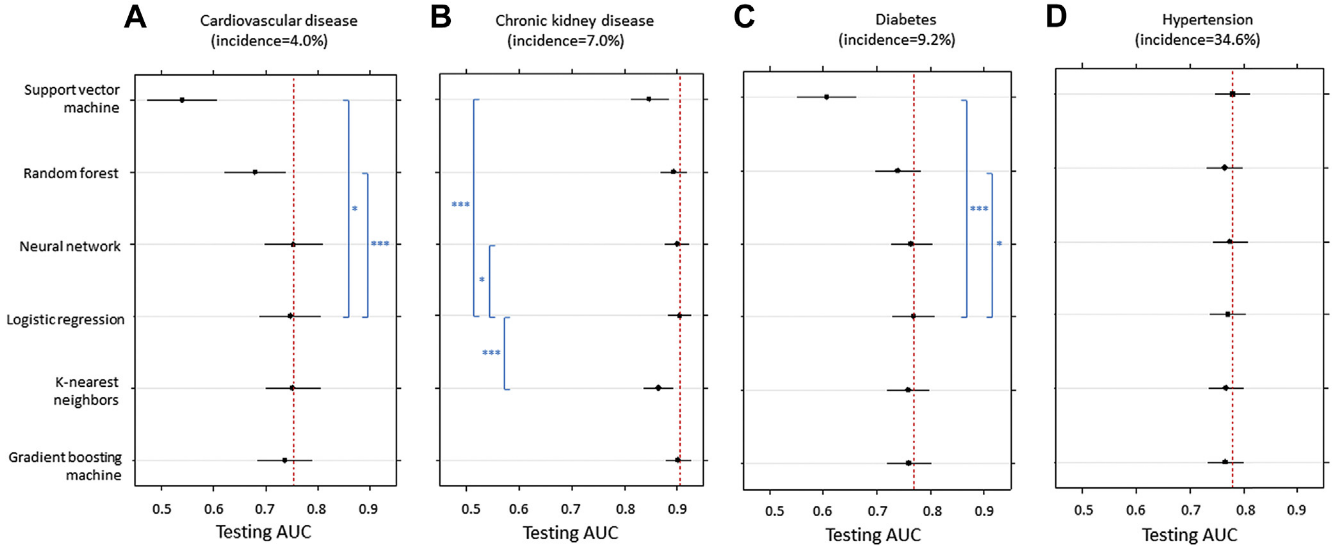

Fig. 2. Performance of the different models considered in the testing data sets for the prediction of A. cardiovascular disease (n = 1,849), B. chronic kidney disease (n = 1,691), C. diabetes (n = 1,497), and D. hypertension (n = 810), expressed in the area under the ROC curve (AUC) with their 95% confidence intervals. The red dotted line corresponds to the best AUC. The AUC corresponded to logistic regression was compared with the AUC from machine learning models using paired Delong's tests with the following notations: * P-value between 0.01 and 0.05; ** P-value between 0.001 and 0.01; *** P-value <0.001. ROC, receiver operating characteristic. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 5. Predictive performance expressed in the area under the ROC curve (AUC), sensitivity (Se), and specificity (Sp) of the six models considered in the prediction of cardiovascular disease, chronic kidney disease, diabetes, and hypertension, in the testing data sets. The bold values corresponded to the highest AUC for each outcome

	Cardiovascular disease			Chronic kidney disease		
	(*n* events = 74, *n* total = 1,849)			(*n* events = 118, *n* total = 1,691)		
Models	AUC + 95% CI	Se	Sp	AUC + 95% CI	Se	Sp
Support vector machine	0.540 [0.47, 0.61]	0.61	0.44	0.848 [0.81, 0.88]	0.86	0.65
Random forest	0.680 [0.62, 0.74]	0.68	0.63	0.895 [0.87, 0.92]	0.81	0.8
Neural network	**0.753 [0.70, 0.81]**	**0.76**	**0.57**	0.901 [0.88, 0.92]	0.84	0.8
Logistic regression	0.748 [0.69, 0.81]	0.74	0.57	**0.905 [0.88, 0.93]**	**0.87**	**0.78**
K-nearest neighbor	0.752 [0.70, 0.8]	0.76	0.6	0.866 [0.84, 0.89]	0.81	0.77
Gradient boosting machine	0.737 [0.68, 0.79]	0.74	0.59	0.903 [0.88, 0.93]	0.86	0.8

Bold values corresponds to the highest AUC for each outcome. *Abbreviations:* CI, confidence interval; ROC, receiver operating characteristic.

0.10 were considered having a significant nonlinear relationship with the outcome. Then, for these predictors, the pattern of the relationship was plotted using generalized additive models to visually inspect the form of the relationship.

2.4.6. *Class imbalance*

Of the four outcomes, three have incidence below 10%, that is, CVD, CKD, and DM. Because the performance of some ML models can be affected by class imbalance, we have performed complementary analyses using a sampling technique, Synthetic Minority Over-sampling Technique (SMOTE), that uses both upsampling and downsampling [42]. To help balance the training set, the algorithm adds new cases via upsampling (by interpolating the existing ones) and downsampling individuals free of disease via random sampling.

3. Results

After exclusion of individuals with missing outcome at baseline or at follow-up, and prevalent cases at baseline, 6,164 (CVD), 5,641 (CKD), 4,992 (DM) and 2,705 (HTN) individuals were included for analyses (Figure 1). The incidences of CVD, CKD, DM, and HTN cases were 4.0% (247/6,164), 7.0% (396/5,641), 9.2% (461/4,992), and 34.6% (936/2,705), respectively. The great majority of the predictors considered were associated to CVD (Table 1), CKD (Table 2), DM (Table 3), and HTN (Table 4). The proportion of missing values was overall low (<5%) except for albumin-to-creatinine ratio (23.1%) and vessel calibers (~15%) (Supplementary materials Tables 1-4).

3.1. *Predictive performance*

The training and testing predictive performances were similar, with mean differences "training - testing" equal to −0.7%, 0.8%, −2.1%, and −1.8% for CVD, CKD, DM, and HTN prediction, respectively (Supplementary Materials Figure 1). There were small differences between training and testing performances; however, the CIs of the testing performances included diagonal, indicating no statistical difference. The CIs were narrower for the training AUC because the bootstrap procedure was repeated 20 times. Figure 2 presents the predictive performance in the testing data set. Logistic regression reached the highest AUC for CKD (0.905 [0.88, 0.93]) and DM (0.768 [0.73, 0.81]) predictions (Table 5). For CVD and HTN prediction models, the best models were neural network (0.753 [0.70, 0.81]) and support vector machine (0.780 [0.747, 0.812]), respectively. However, the differences with logistic regression were small and nonsignificant (ΔAUC = 0.5%, *P*-value = 0.31 for CVD and ΔAUC = 0.9%, *P*-value = 0.10 for HTN). Overall, logistic regression, neural network, and gradient boosting machine were systematically ranked among the best models.

3.2. *Variable importance*

Logistic regression was used to estimate the variable importance of each predictor and was calculated as the absolute value of the z-statistic [33]. The variable importance was scaled so that the sum adds to 100, with higher numbers indicating stronger influence on the response. The three most important variables were ethnicity (being Malay or Indian compared with Chinese) and HbA1c for CVD risk prediction; eGFR, age, and albumin-to-creatinine ratio for CKD risk prediction; blood glucose level, BMI, and family history of DM for DM risk prediction; and systolic blood pressure, age, and HbA1c for HTN risk prediction (Figure 3).

3.3. *Interactions among predictors*

Figure 4 shows the relationship between the training AUC and the maximum interaction depth, in accordance with the number of trees. For low interaction depth, the best AUC

Diabetes			Hypertension		
(n events = 138, n total = 1,497)			(n events = 280, n total = 810)		
AUC + 95% CI	Se	Sp	AUC + 95% CI	Se	Sp
0.606 [0.55, 0.66]	0.64	0.48	**0.780 [0.747, 0.812]**	**0.85**	**0.6**
0.739 [0.70, 0.78]	0.72	0.64	0.765 [0.731, 0.799]	0.8	0.63
0.764 [0.72, 0.80]	0.78	0.62	0.775 [0.742, 0.808]	0.83	0.58
0.768 [0.73, 0.81]	**0.74**	**0.63**	0.770 [0.737, 0.803]	0.8	0.6
0.758 [0.72, 0.80]	0.82	0.58	0.768 [0.735, 0.801]	0.81	0.61
0.760 [0.72, 0.80]	0.67	0.68	0.767 [0.734, 0.801]	0.84	0.56

corresponded to the highest number of trees, and reciprocally, for high interaction depth, the best AUC corresponded to the lowest number of trees. In other words, the model can learn faster from the data if more interactions are allowed. For CVD, DM, and HTN, the best performance corresponded to very low interaction depth (1 or 2). For CKD, there was a great improvement from interaction equal to two, followed by a plateau with AUC around 0.905.

3.4. Nonlinearity

Among the predictors, the following had nonlinear relationships with the outcome: systolic blood pressure with CVD, age and albuminuria with CKD, systolic blood pressure and total blood glucose with DM, and HbA1c with HTN. Despite the significance of the nonlinear effects, the visual inspection of the form of the relationships showed very small deviances to linearity for all these predictors (Figure 5).

3.5. Class imbalance

The results of the performance when using SMOTE showed very similar results except for the support vector machine model (Supplementary Materials Figure 1). For this model, the performance was greatly improved with AUC increasing from 0.54 to 0.73 for CVD, from 0.85 to 0.91 for CKD, and 0.61 to 0.74 for DM. There was no improvement for all the models after sampling for HTN. As for the results without SMOTE, no ML model outperforms logistic regression. To confirm these results, we have also performed a cost-sensitive training for the support vector machine model by imposing heavier cost when errors are made in the case class (inversely proportional to the class frequency, ie, 24 times more weight on the case class) and found the same improvement (AUC with cost-sensitive training = 0.75).

4. Discussion

We showed that logistic regression yields as good performance as ML models to predict the risk of major chronic diseases in an epidemiological study of moderate sample size typical of many studies with a limited number of incident events and a limited set of simple clinical predictors. Among the different models, logistic regression had the best performance for CKD and DM risk prediction. For

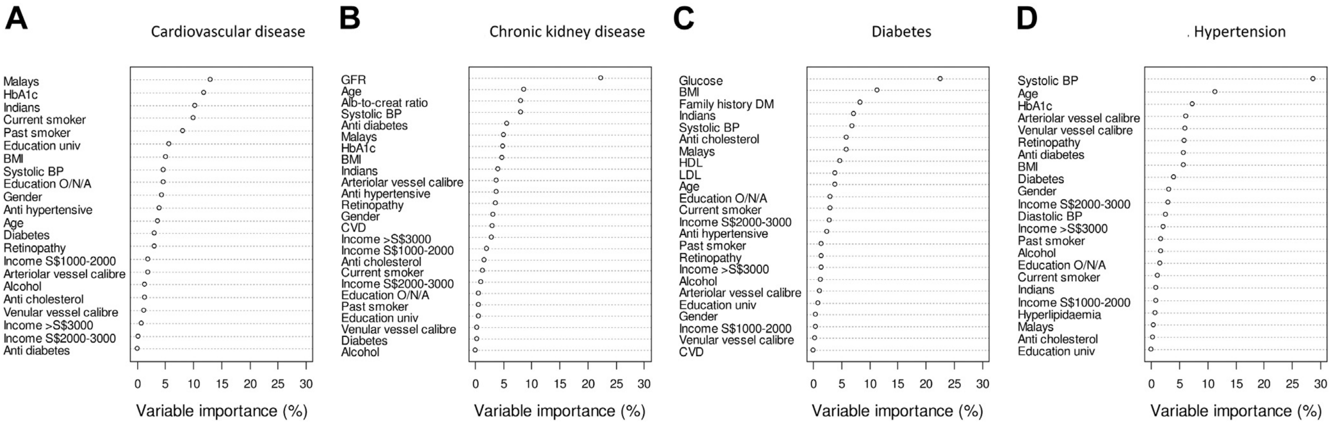

Fig. 3. Variable importance of the predictors for the predictive models for A. cardiovascular disease, B. chronic kidney disease, C. diabetes, and D. hypertension, using logistic regression. The variable importance was calculated using the absolute z-statistic of each predictor. BMI, body mass index; BP, blood pressure; CVD, cardiovascular disease; DM, diabetes mellitus; GFR, glomerular filtration rate; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

CVD and HTN, logistic regression was ranked third, but the difference in AUC with the best models (neural network and support vector machine) was small and nonsignificant. Logistic regression, gradient boosting machine, and neural network were systematically ranked among the best models. The training and testing predictive performances were similar, suggesting a low risk of overfitting. The most influent variables were consistent with knowledge on these metabolic diseases. Given the risk of overfitting and the lack of interpretability of some ML models (eg, neural network), we therefore suggest that traditional regression models should continue to have a key role in disease risk prediction when using a limited set of simple clinical predictors. The use of ML techniques may not be warranted in such studies.

The overall predictive performances reached for each outcome (AUC = 0.75 for CVD, AUC = 0.90 for CKD, AUC = 0.77 for DM, and AUC = 0.78 for HTN) were consistent with those in previous studies [11,23,43] Compared with other models, logistic regression was performing very well to predict the risk of common chronic disease over a period of ~6 years. In accordance with previous studies, when the number of cases is limited and when considering a small number of simple clinical predictors, logistic regression is an easy-to-use and appropriate model for disease risk prediction [12,13,16,17]. As suggested by previous studies, advanced ML techniques should be used in highly complex medical prediction problems (strong interactions and nonlinearities) if very large data sets are available [18,44] or when simple methods cannot be used because of the nature of the data, such as images. Support vector machine was strongly affected by class imbalance, as previously described [45]. Using a method to help balance the data (either weighting or sampling), the AUC increased up to 20% for this model. For HTN risk prediction, where the classes were more balanced (incidence around 35%), all the models were performing similarly well and support vector machine had the highest performance. When using a limited set of simple clinical predictors, we suggest using either logistic regression or an ML model interpretable and not sensitive to class imbalance, such as random forest or gradient boosting machine. We believe that neural network and support vector machine, less interpretable, should be used for more complex data and when large data sets are available (and using appropriate methods to handle class imbalance is needed).

Furthermore, the good performance of logistic regression compared with ML models is due to (1) the overall small amount of interaction among the predictors and (2) the absence of strong nonlinear relationships between predictors and the risk of diseases. Regarding the interactions, their overall amount between predictors was very low for CVD-, DM-, and HTN-predictive risk models. That amount was larger for the CKD model; however, the performance improved only very slightly when considering interactions higher than order 2. One main interaction was identified using gradient boosting machine between the two most influent predictors: glomerular filtration rate and age. Surprisingly, this interaction (not included in the initial model) was not significant in the logistic regression model, suggesting that tree-based ML methods might be more sensitive to interaction than logistic regression. Regarding the linearity, few relationships were statistically nonlinear (with a significant nonlinear term); however, the deviances to linearity were visually very small, meaning that the implicit linear assumption made in the logistic regression was overall valid.

Several studies have found the logistic regression to perform very well for prediction with predictive performance, at least as good as ML techniques [12−17]. Moreover, in some studies that claimed the better performance of ML, the corresponding increment was limited (increment in AUC between 1 and 3%) [4,8,10,11]. It has been already suggested that simple methods typically yield performance almost as good, or even superior, as more sophisticated methods [9]. For example, in a study aiming at comparing

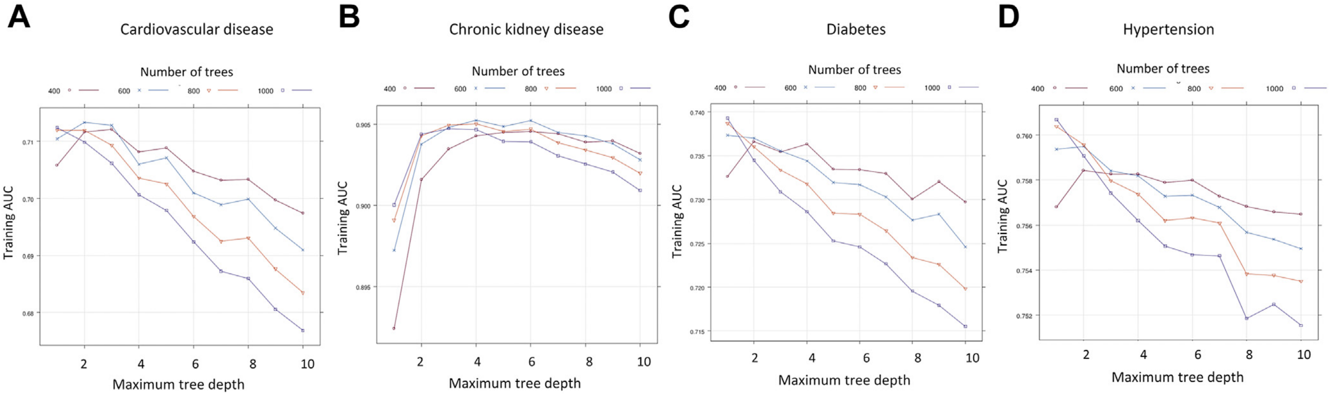

Fig. 4. Predictive performance according to the maximum interaction depth and the number of trees using gradient boosting machine, expressed in area under the ROC curve (AUC) for A. cardiovascular disease, B. chronic kidney disease, C. diabetes, and D. hypertension. ROC, receiver operating characteristic. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

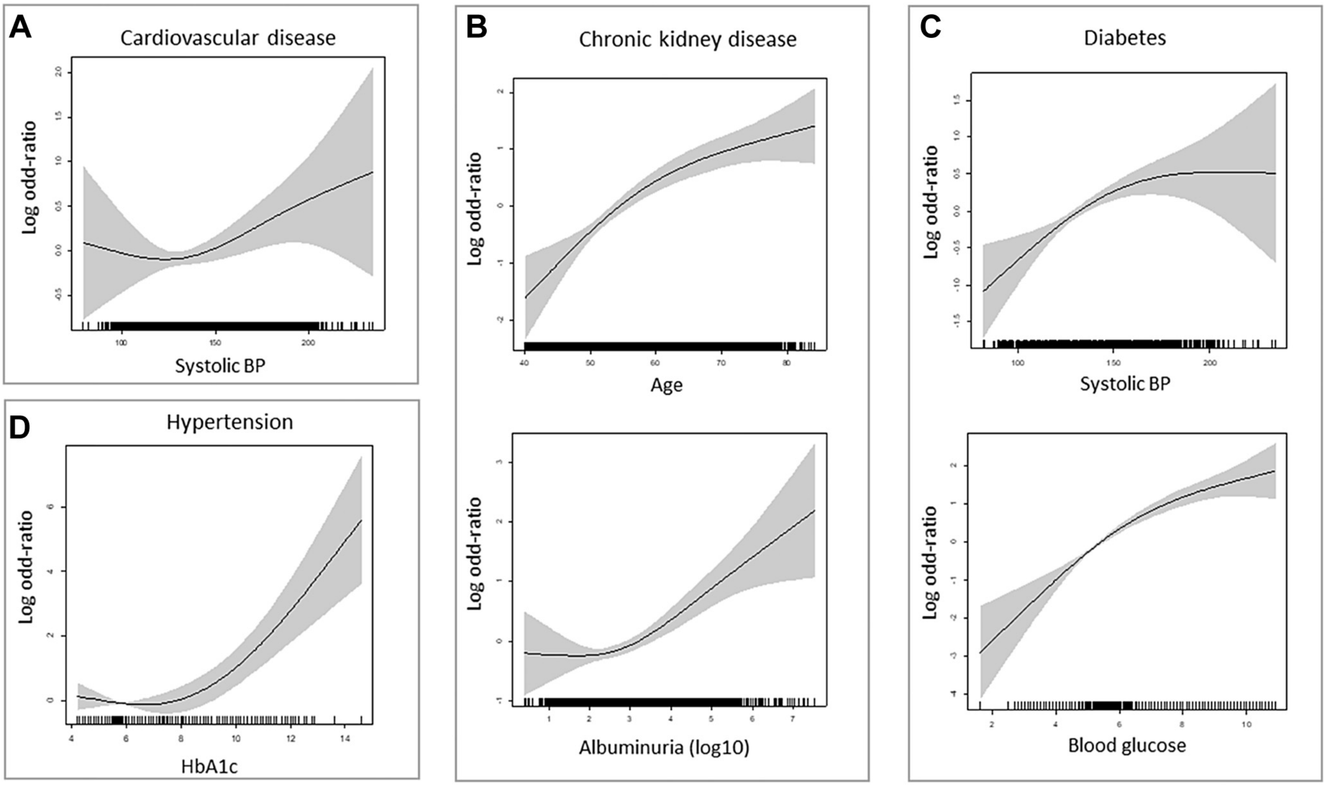

Fig. 5. Relationships between predictors that have a nonlinear effect on the risk of A. cardiovascular disease, B. chronic kidney disease, C. diabetes, and D. hypertension, expressed in log odd ratios with their 95% confidence intervals (using generalized additive model). Nonlinear effects were tested by adding squared terms of the continuous predictors in a logistic regression model. BP, blood pressure; HbA1c, glycosylated hemoglobin.

22 decision trees, 9 statistical, and 2 neural networks on 32 data sets, the old algorithm linear discriminant analysis has a mean error rate close to the best and logistic regression second [46].

This study has the following strengths. First, the analyses included at baseline about 10,000 participants and between 5,000 and 6,000 individuals for each incident outcome. This is typical of many epidemiological studies of chronic diseases [23,47,48]. Second, we used data on four common outcomes to explore different scenarios and check the robustness of the results in accordance with the incidence of the outcome. Third, interaction and nonlinearities were investigated using appropriate analytical tools and these characteristics of the data were put in parallel with the performance of the different models. Finally, we have estimated both training and testing predictive performance and showed the absence of overfitting issue. This is an important point because ML models can easily overfit.

The following limitations should also be acknowledged. First, only a limited number of predictors were considered. We have restricted our analyses to predictive modeling with known or possible risk factors. Our conclusion cannot be generalized to data sets with more predictors. Second, all the four data sets showed low amount of interaction and relationships very close to linearity. It would be interesting to explore the performance of the models in the presence of more interactions and more nonlinearities. However, logistic regression might still perform well if some refinements are considered, for example, by including interaction terms or splines or polynomial terms into account for eventual nonlinear effects. Finally, CVD was self-reported. However, the incidence was similar to the 6-year incidence reported in the Framingham cohort [23], the reliability has been checked, and this definition has been used in several articles [28,49]. Moreover, we are interested here in the relative performance of different models rather than the absolute performance of a single predictive model.

5. Conclusion

In conclusion, we demonstrate that logistic regression yields as good performance as ML models to predict the risk of common chronic diseases over ~6 years in an epidemiological study (1) of moderate sample size typical of many studies, (2) with a limited number of events, (3) with a limited set of simple clinical predictors, and (4) using some of the most commonly used ML models. We suggest that traditional regression models should continue to have a key role in disease risk prediction. Further studies are needed to confirm this result for different settings and study characteristics.

CRediT authorship contribution statement

Simon Nusinovici: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing - original draft. **Yih Chung Tham:** Methodology, Validation, Writing - review & editing. **Marco Yu Chak Yan:** Methodology, Software, Writing - review & editing. **Jialiang Li:** Formal analysis, Software, Writing - review & editing. **Charumathi Sabanayagam:** Supervision, Funding acquisition, Writing - review & editing. **Tien Yin Wong:** Supervision, Funding acquisition, Writing - review & editing. **Ching-Yu Cheng:** Supervision, Validation, Writing - review & editing.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.03.002.

References

[1] 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.

[2] Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, et al. Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg 2018;109:476−486.e1.

[3] Kruppa J, Liu Y, Diener H-C, Holste T, Weimar C, König IR, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: applications. Biom J 2014;56:564−83.

[4] Singal AG, Mukherjee A, Elmunzer BJ, Higgins PDR, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol 2013;108:1723−30.

[5] Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of in-hospital mortality in emergency department patients with Sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 2016;23:269−78.

[6] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med 2016;44:368−74.

[7] Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017;38:1805−14.

[8] Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. J Investig Med 1995;43:468−76.

[9] Hand DJ. Classifier technology and the illusion of progress. Stat Sci 2006;21:1−14.

[10] Taylor RA, Moore CL, Cheung K-H, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. PLoS One 2018;13:e0194085.

[11] Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y, et al. Screening for prediabetes using machine learning models. Comput Math Methods Med 2014;2014:618976.

[12] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12−22.

[13] Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol 2013;66:398−407.

[14] König IR, Malley JD, Weimar C, Diener H-C, Ziegler A. German Stroke Study Collaboration. Practical experiences on the necessity of external validation. Stat Med 2007;26:5499−511.

[15] König IR, Malley JD, Pajevic S, Weimar C, Diener H-C, Ziegler A. Patient-centered yes/no prognosis using learning machines. Int J Data Min Bioinform 2008;2:289−341.

[16] van der Ploeg T, Smits M, Dippel DW, Hunink M, Steyerberg EW. Prediction of intracranial findings on CT-scans by alternative modelling techniques. BMC Med Res Methodol 2011;11:143.

[17] Van Calster B, Valentin L, Van Holsbeke C, Testa AC, Bourne T, Van Huffel S, et al. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. BMC Med Res Methodol 2010;10:96.

[18] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137.

[19] El-Solh AA, Hsiao C-B, Goodnough S, Serghani J, Grant BJB. Predicting active pulmonary tuberculosis using an artificial neural network. Chest 1999;116:968−73.

[20] Ward MM, Pajevic S, Dreyfuss J, Malley JD. Short-term prediction of mortality in patients with systemic lupus erythematosus: classification of outcomes using random forests. Arthritis Rheum 2006;55:74−80.

[21] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 2010;10:16.

[22] Meguid El Nahas A, Bello AK. Chronic kidney disease: the global challenge. Lancet 2005;365:331−40.

[23] D'Agostino RB, Pencina MJ, Massaro JM, Coady S. Cardiovascular disease risk assessment: insights from Framingham. Glob Heart 2013;8:11−23.

[24] Rosman M, Zheng Y, Wong W, Lamoureux E, Saw SM, Tay WT, et al. Singapore Malay Eye Study: rationale and methodology of 6-year follow-up study (SiMES-2). Clin Exp Ophthalmol 2012;40:557−68.

[25] Sabanayagam C, Yip W, Gupta P, Mohd Abdul RBB, Lamoureux E, Kumari N, et al. Singapore Indian Eye Study-2: methodology and impact of migration on systemic and eye outcomes. Clin Exp Ophthalmol 2017;45:779−89.

[26] Lavanya R, Jeganathan VSE, Zheng Y, Raju P, Cheung N, Tai ES, et al. Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. Ophthalmic Epidemiol 2009;16:325−36.

[27] Foong AWP, Saw S-M, Loo J-L, Shen S, Loon S-C, Rosman M, et al. Rationale and methodology for a population-based study of eye diseases in Malay people: the Singapore Malay eye study (SiMES). Ophthalmic Epidemiol 2007;14:25−35.

[28] Wong MYZ, Man REK, Gupta P, Lim SH, Lim B, Tham Y-C, et al. Is corneal arcus independently associated with incident cardiovascular disease in asians? Am J Ophthalmol 2017;183:99−106.

[29] Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009;150:604−12.

[30] Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. Lancet 2017;389:1238−52.

[31] Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016i2416.

[32] Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011;343: d7163.

[33] Kuhn Max. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.

[34] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Br J Surg 2015; 102:148–58.

[35] Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9: 1–12.

[36] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[37] Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28: 112–8.

[38] Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 2013;3:e002847.

[39] Kuhn M, Johnson K. Over-Fitting and Model Tuning. Appl. Predict. Model.. New York, NY: Springer New York; 2013: 61–92.

[40] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5.

[41] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.

[42] Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[43] Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk models to predict hypertension: a systematic review. PLoS One 2013;8:e67370.

[44] Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. Biom J 2014;56:601–6.

[45] Batuwita R, Palade V. Class imbalance learning methods for support vector machines. Imbalanced Learn. Hoboken, New Jersey: Found. Algorithms Appl. Wiley; 2013.

[46] Lim T-S, Loh W-Y, Shih Y-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Learn 2000;40:203–28.

[47] Bild DE. Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol 2002;156:871–81.

[48] Hofman A, Breteler MMB, van Duijn CM, Janssen HLA, Krestin GP, Kuipers EJ, et al. The Rotterdam Study: 2010 objectives and design update. Eur J Epidemiol 2009;24:553–72.

[49] Gupta P, Gan ATL, Man REK, Fenwick EK, Tham Y-C, Sabanayagam C, et al. Risk of incident cardiovascular disease and cardiovascular risk factors in first and second-generation Indians: the Singapore Indian eye study. Sci Rep 2018;8:14805.