JBE Journal of Biostatistics and Epidemiology
J Biostat Epidemiol. 2025;11(3): 335-345

ORIGINAL ARTICLE
DOI: *https://doi.org/10.18502/jbe.v11i3.21373/*
*Received **May** 15, 2025; Accepted **June** 30, 2025;*
*Published online **September** 15, 2025.*

Analyzing Cardiovascular Disease Risk Factors Using Generalized Logistic Logic Regression: A Retrospective Study

Habibollah **Esmaily**[1,2], Mahbubeh **Jahani**[3], Mohammad Ali **Kianfard**[1,4*] , Majid **Ghayour-Mobarhan**[5]

[1]Student Research Committee, Mashhad University of Medical Sciences, Mashhad, Iran.
[2]Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.
[3]Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[4]Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.
[5]International UNESCO Center for Health-Related Basic Sciences and Human Nutrition, Mashhad University of Medical Sciences, Mashhad, Iran.

ABSTRACT

Introduction: Cardiovascular disease (CVD) is a general term that refers to diseases of the heart or blood vessels. Logic regression is a machine learning method that is commonly used when the number of predictor variables is high, and it can account for interaction effects between predictor variables. As CVD can be influenced by multiple factors, this study was conducted to identify variables related to CVD and predict the occurrence of CVD using generalized logistic logic regression.

Methods: The present study is a retrospective study utilizing data from phase one of the MASHAD study. The analysis was performed on the information of 7,385 individuals. Generalized logistic logic regression analysis was performed using the "LogicReg" package in R software.

Results: Out of the 7385 individuals included in this study, 235 (3.2%) were diagnosed with CVD, while 7150 (96.8%) did not have CVD. Of the variables examined, age, anxiety, depression, metabolic syndrome, and family history were significant as main effects, and an interaction between smoking status and education had a significant effect.

Conclusion: Based on the findings of this study, it can be tentatively concluded that for CVD, the existence of interaction effects among the mentioned risk factors may not be a significant concern. In other words, the primary effects of each variable may be more important, as these variables appear to play a role in CVD independently of each other.

Key words: Generalized logistic logic regression; Cardiovascular disease; Machine learning; Interaction effects; Risk factors

*Corresponding Author:
kianfardma@gmail.com

INTRODUCTION

In the opinion of the World Health Organization (WHO), cardiovascular disease (CVD) is universally the leading cause of death. In 2019, an evaluated 523 million individuals worldwide were affected by CVD, resulting in 18.6 million deaths. In 2020, CVD was responsible for approximately 19.1 million deaths worldwide.[1, 2]

It is estimated that the number of disability-adjusted life years (DALYs) attributed to CVD, including coronary artery disease (CAD), in Iranian adults aged over 30 years will increase by more than twofold from 847,309 DALYs in 2005 to 1,728,836 DALYs in 2025, according to projections spanning the period from 2005 to 2025.[3, 4]

CVD is widely recognized to have a complex and multifactorial etiology. Therefore, thorough examination of CVD risk factors and primary diagnosis of individuals at high risk of CVD are crucial for timely intervention and constitutes vital component of healthcare policies in many countries.[5]

Regression models are conventional method for analyzing data and examining relationships between predictor variables and response variable. They predict the response variable.[6] In practice, predictor variables often exhibit interaction effects, whereby the effect of a given variable differs depending on the presence or absence of another variable. However, identifying and incorporating these interaction effects can increase a regression model's complexity. As a result, two-way and three-way interaction effects are typically employed in the final model.[7]

When the number of predictor variables is large, particularly with dichotomous variables, higher-order interactions can affect the response variable's fit. To address this, combination variables created from predictors can be used as new independent variables rather than including all original variables in model fitting. These constructed variables are called Boolean combinations.[6]

Logic regression, introduced by Ruczinski, is a machine learning technique that serves as a generalized regression and classification method. In this method, independent variables are constructed as Boolean combinations of binary variables, and generalized logic regression includes quantitative variables in addition to binary independent variables. Generalized logic regression identifies optimal Boolean combinations of binary and quantitative variables that, when used as predictors, provide the best fit for the response variable. These Boolean combinations are linked to the response variable via a link function. A lower score function of the regression model indicates better fit.[6]

Logic regression offers several advantages over other methods, including its regression form, which allows for straightforward interpretation of coefficients, the ability to include interactions between multiple variables in the form of a Boolean expression, and the ability to summarize variables.[8] Given CVD's multifactorial nature, this study uses generalized logistic logic regression to identify associated factors and predict occurrence.

METHODS

The present study is a retrospective study utilizing data from phase one of the MASHAD study. The MASHAD study is a cross-sectional investigation conducted from 1 January 2010 to 1 January 2011. The study sample comprised all individuals registered in the MASHAD study database. The initial sample size was 9,884 participants; after removing cases with missing or incomplete data, the final analysis included 7,385 participants.

The variables considered in this study included age, body mass index (BMI), physical exercise, cholesterol, high sensitivity C-reactive protein (hsCRP), low-density lipoprotein cholesterol (LDL-C), marital status, smoking status, gender, employment status, family history, metabolic syndrome, anxiety, depression, and education.

To perform generalized logistic logic regression, the study variables must be either quantitative or dichotomous. The study included six quantitative variables, including physical exercise, age, BMI, cholesterol, LDL, and hsCRP, and six intrinsically dichotomous variables: smoking status (nonsmoker/smoker), marital status (single/married), gender (male/female), employment status (unemployed/employed), family history (no/yes), and metabolic syndrome (no/yes). Metabolic syndrome was defined by six components [fasting blood glucose (FBG), systolic blood pressure (SBP), diastolic blood pressure (DBP), triglycerides (TG), waist circumference, and high-density lipoprotein cholesterol (HDL-C)]. If at least three of these variables exceeded the cutoff point set by the WHO, the person was considered to have metabolic syndrome. These variables were entered into the model in their original form.

However, three variables were not dichotomous and were transformed into dichotomous variables to be included in the model. Anxiety was assessed using the Beck Anxiety Inventory (BAI), with scores ranging from 0 to 63. The original four categories (0-7, 8-15, 16-25, and 26-63) were dichotomized into low anxiety (0-15) and high anxiety (16-63). Depression was measured using the Beck Depression Inventory-II (BDI-II), with four original categories: minimal (0-13), mild (14-19), moderate (20-28), and severe (29-63). These were dichotomized into low depression (0-19) and high depression (20-63). Education was originally categorized as low (illiterate to high school), medium (bachelor's), and high (master's/doctorate). Categories were merged to create two groups: less than diploma and more than diploma.

The dataset exhibited a significant class imbalance, with only 3.2% of individuals diagnosed with CVD (235 cases) compared to 96.8% without CVD (7,150 cases). Given machine learning methods' sensitivity to class imbalance, we addressed this issue using resampling methods before model fitting. To address this imbalance and mitigate its potential bias on model performance, we employed the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTE-NC). SMOTE-NC was chosen because it effectively handles mixed data types (categorical and continuous variables) by generating synthetic samples for the minority class (CVD patients) while preserving the underlying data distribution. This method creates new instances by interpolating between existing

minority class samples and their nearest neighbors, thus avoiding overfitting associated with simple random oversampling.

To prevent observer error, machine learning models use separate training and test sets. We split the data into training (75%; 176 CVD and 5,362 healthy) and test (25%; 59 CVD and 2,145 healthy) sets. The model was trained on the training set, with parameters estimated, while the test set evaluated its performance and accuracy.

The SMOTE-NC algorithm was applied exclusively to the training set to balance class distribution while preserving the original test set for evaluating model generalizability. To mitigate overfitting, synthetic CVD cases were generated until the minority class reached half the size of the majority class (2,681 CVD vs. 5,362 healthy).

Following class balancing, a simulated annealing search algorithm identified optimal Boolean combinations (logic trees) for inclusion in the logistic logic regression. Ten-fold cross-validation was implemented during model selection to prevent overfitting. The final model consisted of 5 Boolean combinations and 6 variables (Figure 1).

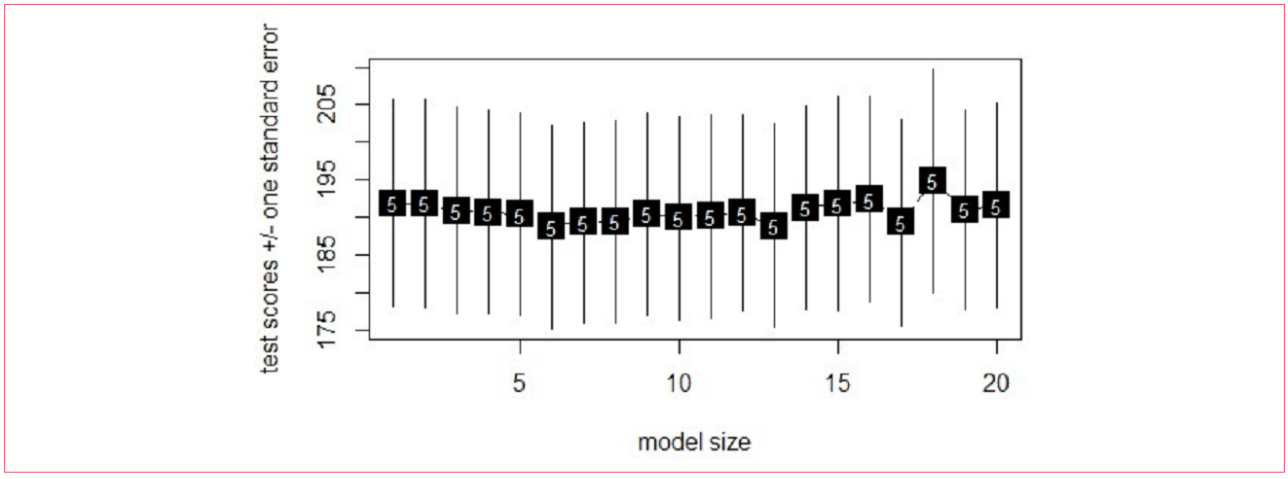

Figure 1. Cross-validation results showing model performance (scores ± standard error) across different model sizes. The x-axis indicates the total number of predictors, with the count of Boolean combinations displayed within each square.

We evaluated generalized logistic logic regression against binary logistic regression (including only main effects) by computing deviation statistics, Akaike Information Criterion (AIC), sensitivity, and specificity. Model performance in predicting CVD incidence was assessed using receiver operating characteristic (ROC) curves and area under the curve (AUC) values.

Analyses were performed in R (v4.1.2) using the LogicReg package for generalized logistic logic regression and the smotefamily package for SMOTE-NC processing. Continuous variables are presented as mean ± standard deviation, and categorical variables as frequency (percentage). Statistical significance was set at $p < 0.05$.

Ethical considerations

All participants provided written informed consent for data use in this study. To ensure confidentiality, each individual was assigned a unique identification code. The study protocol received ethical approval from the Mashhad University of Medical Sciences Health Faculty Ethics Committee (Approval Code: IR.MUMS.FHMPM.REC.1401.024).

RESULTS

The study included 7,385 participants, with 235 (3.2%) diagnosed with CVD and 7,150 (96.8%) without CVD. Tables 1 and 2 present the demographic characteristics for quantitative and qualitative variables, respectively, stratified by CVD status.

Table 1. Sociodemographic characteristics for numerical variables in all patients, with CVD and No CVD

Variables	CVD	No CVD	Total
	Mean±SD	Mean±SD	Mean±SD
Physical exercise (Met-h/day)	1.53±0.28	1.58±0.28	1.58±0.28
Age (year)	54.31±6.85	48.29±7.74	48.48±7.8
BMI (kg/m^2)	28.83±4.56	27.76±4.67	27.8±4.6
Cholesterol (mg/dl)	199.85±41.56	190.97±37.89	191.25±38.04
LDL (mg/dl)	122.30±35.50	116.51±34.69	116.7±34.73
HsCRP (mg/L)	4.83±7.92	3.78±7.79	3.8±7.8

SD, Standard Deviation; CVD, Cardiovascular disease; BMI, Body Mass Index; LDL, Low-Density Lipoprotein; HsCRP, High sensitivity C-Reactive Protein

Table 2. Sociodemographic characteristics for categorical variables in all patients, with CVD and No CVD

Variables	Items	CVD	No CVD	Total
		Mean±SD	Mean±SD	Mean±SD
Smoking status	Smoker	94 (4.1)	2205 (95.9)	2299 (31.1)
	Nonsmoker	141 (2.8)	4945 (97.2)	5086 (68.9)
Family History	Yes	104 (4.1)	2422 (95.9)	2526 (34.2)
	No	131 (2.7)	4728 (97.3)	4859 (65.8)
Metabolic Syndrome	Yes	128 (5.9)	2054 (94.1)	2182 (29.5)
	No	107 (2.1)	5096 (97.9)	5203 (70.5)
Anxiety	Low	162 (2.8)	5608 (97.2)	5770 (78.1)
	High	73 (4.5)	1542 (95.5)	1615 (21.9)
Depression	Low	173 (2.9)	5851 (97.1)	6024 (81.6)
	High	62 (4.6)	1299 (95.4)	1361 (18.4)
Education	Less than diploma	154 (3.9)	3781 (96.1)	3935 (53.3)
	More than diploma	81 (2.3)	3369 (97.7)	3450 (46.7)
Marital status	Single	18 (4)	432 (96)	450 (6.1)
	Married	217 (3.1)	6718 (96.9)	6935 (93.9)
Gender	Male	111 (3.4)	3143 (96.6)	3254 (44.1)

	Female	124 (3)	4007 (97)	4131 (55.9)
Employment status	Unemployed	70 (2.4)	2877 (97.6)	2947 (39.9)
	Employed	165 (3.7)	4273 (96.3)	4438 (60.1)

SD, Standard Deviation; CVD, Cardiovascular disease; BMI, Body Mass Index; LDL, Low-Density Lipoprotein; HsCRP, High sensitivity C-Reactive Protein

Table 3 shows the optimal predictor subset from cross-validation, comprising 6 variables and 5 Boolean combinations. Generalized logistic logic regression revealed significant associations ($p < 0.05$) between CVD and age, anxiety, metabolic syndrome, family history, depression, and smoking status × education (interaction effect).

The analysis revealed a significant positive association between age and CVD risk, with each additional year increasing the risk by 1.1-fold (OR = 1.1), corresponding to a 10% elevation in annual probability. For categorical variables, nonsmokers with more than diploma education demonstrated a 50% reduced CVD risk (OR = 0.5) compared to smokers with less than diploma educational attainment. Similarly protective associations were observed for: individuals with low anxiety levels with 30% lower risk (OR = 0.7), those without metabolic syndrome with 52% lower risk (OR = 0.48), and participants with no family history of CVD with 40% lower risk (OR = 0.6). Conversely, participants with high depression levels exhibited 40% greater CVD risk (OR = 1.4) than those with lower depression levels.

Table 3. The result of the generalized logistic logic regression to recognize factors related with CVD

Variables		Items	OR[a]	95% CI[b] for OR	P-V
Quantitative variables	1	Physical exercise (Met-h/day)	0.4	(0.326 , 0.681)	0.681
	2	Age (year)	1.095	(1.075 , 1.116)	0.000
	3	BMI (kg/m^2)	1.005	(0.975 , 1.037)	0.739
	4	Cholesterol (mg/dl)	0.999	(0.993 , 1.005)	0.735
	5	LDL (mg/dl)	1.002	(0.996 , 1.009)	0.527
	6	HsCRP (mg/L)	1.005	(0.990 , 1.019)	0.545
Boolean combinations	1	Education (Less than diploma*) and Smoking (smoker*)	0.522	(0.369 , 0.738)	0.000
	2	Anxiety (High*)	0.703	(0.512 , 0.967)	0.030
	3	Metabolic Syndrome (Yes*)	0.483	(0.359 , 0.649)	0.000
	4	Family History (Yes*)	0.598	(0.457 , 0.784)	0.000
	5	Depression (Low*)	1.437	(1.031 , 2.003)	0.032

[a]Odds Ratio; [b]Confidence Interval; *Reference Category;
BMI, Body mass index; LDL, Low-density lipoprotein; hsCRP, High sensitivity C-Reactive protein

We conducted binary logistic regression analysis including only main effects. The results almost paralleled those of generalized logistic logic regression, identifying physical exercise, age, smoking status, family history, metabolic syndrome, anxiety, and depression as significant predictors of CVD risk (detailed results not shown). Model comparison revealed superior performance of generalized logistic logic regression, evidenced by lower AIC values (1881 vs. 1889) and deviation statistics (1857

vs. 1863) relative to binary logistic regression. This demonstrates enhanced predictive capability of the generalized logistic logic regression approach for CVD risk assessment. Additional performance metrics are presented in Table 4.

Table 4. Comparing the efficiency of the binary logistic regression and the generalized logistic logic regression

Indices	Binary logistic regression	Generalized logistic logic regression
Deviation statistic	1863	1857
Akaike information criterion	1889	1881
Sensitivity	0.75	0.79
Specificity	0.87	0.90
AUC*	0.77	0.80

* Area under the ROC curve

DISCUSSION

Identifying factors associated with diseases and cancers and analyzing their interaction effects to predict disease risk is a fundamental task in medical sciences. To achieve this, a method that can identify risk factors and analyze their interaction effects on CVD is necessary. Various sources suggest that risk factors for CVD include high blood pressure, advanced age, psychosocial factors, obesity, lack of regular physical activity, smoking, dyslipidemia, high FBG, fibrinogen, lipoprotein(a), diabetes, homocysteine, acute phase protein, sex, blood lipid disorders (high cholesterol, high TG, low HDL and high LDL), alcohol consumption, inappropriate diet, race, and family history.[9-14]

This article aims to identify factors associated with CVD using generalized logistic logic regression, which is a machine learning method. Given that CVD can be influenced by multiple factors, identifying the factors and their interaction effects can provide valuable insights. The results of this study indicate that age, less than diploma education, smoking, high stress, metabolic syndrome, family history, and high depression have a significant impact on the probability of developing CVD. Among the variables examined, the only significant interaction effect was observed between education and smoking. Accounting for this interaction reduced the deviation statistic from 1863 for the binary logistic model, which resulted from the main effects alone, to 1857 for the generalized logistic logic regression. This improvement in model fit suggests an increase in the predictive power of the model.

For instance, Weng et al. studied the accuracy of machine learning models in predicting the risk of CVD, including gradient boosting machines and random forest models. The study identified age, smoking, HDL-C, TG, hemoglobin A1c (HbA1c), SBP, sex, BMI, and cholesterol as the 9 variables that significantly affect the risk of CVD.[15]

Our study shares similarities with the study of Weng et al, with the exception of the variables of race and socioeconomic status (SES), which were not included in our investigation and were not discussed in our research. Our results were also different from theirs in terms of the significance of sex, cholesterol, and BMI.

Hongzong et al conducted a study to evaluate the precision and efficiency of the support vector machine algorithm in recognizing and distinguishing between individuals with CAD and healthy individuals. Their findings suggested that out of the many variables investigated, only six variables (cholesterol, HDL-C, TG, LDL-C, age, and FBG) were chosen as variables affecting the classification of observations.[16]

Although our study did not find LDL-C and cholesterol to have a significant effect, the results of Hongzong et al's study were similar to ours. Specifically, the gender variable was not significant in Hongzong et al's study, which is consistent with our findings. However, the variables of family history and smoking, which were significant in our study, were not significant in Hongzong's study.

The WHO identifies unhealthy diet, physical inactivity, tobacco use, and harmful alcohol consumption as the most important behavioral risk factors for heart disease. These risk factors can lead to increased blood pressure, increased blood glucose, increased blood lipids, overweight, and obesity. Additionally, poverty, stress, and hereditary factors are other determinants of CVDs.

Our study findings support the WHO's statements, except for the variables that were not considered in our investigation, such as alcohol consumption, hereditary factors, and unhealthy diet.

Several studies have identified hypertension, diabetes mellitus, dyslipidemia, obesity, smoking, and age as the most frequent risk factors for the occurrence of cardiovascular diseases. These factors are also involved in the development and progression of atherosclerosis.[17-18]

Strengths of study

To the best of our knowledge, this is the first study to apply generalized logistic logic regression to CVD patients, which is one of the unique aspects of our study. Additionally, the participation rate in our investigation was high. Our study's use of the generalized logistic regression method has yielded intriguing results concerning the importance of different risk factors. These findings may inform further exploration of diverse predictive risk factors and the development of new risk prediction approaches.

An important advantage of logistic regression over some models, such as neural networks, is that it is fully interpretable. The coefficients can be interpreted, and the model can be evaluated using statistics related to the type of regression used. Additionally, interactions between multiple variables can be included in the form of a Boolean expression. Using the significant variables identified in our study, including age, smoking, education, family history, metabolic syndrome, stress, and depression, it is possible to predict the probability of CVD with ease using this model.

Limitations of the study

This study has some limitations that should be considered. One important limitation is the lack of data regarding the duration and quantity of cigarette smoking, which may be related to CAD occurrence.

Additionally, accurate data on alcohol consumption were not available due to legal restrictions and cultural norms, although this is likely not a significant issue, as the majority of the population are Muslims who do not consume alcohol.

The failure to measure some important variables, such as diet and socioeconomic status, is another limitation of this study. Furthermore, the sample population was primarily urban, which may limit the generalizability of our findings to rural populations.

Future implications

This study was conducted on a large population of CVD patients in Mashhad. The use of machine learning methods, particularly the generalized logistic regression method, and the use of routine clinical data available within electronic records in multiple countries suggest the potential for the applicability of our findings to other populations and health systems.

Machine learning approaches offer an exciting opportunity to improve and individualize CVD risk assessment. This may aid in the move toward personalized medicine, allowing for more tailored risk management to individual patients.[19-20] The improvement in predictive accuracy observed in this study should be further investigated using machine learning with other large clinical data sets in diverse populations and to predict other disease outcomes.

CONCLUSION

Based on the findings of this study, it can be tentatively concluded that for CVD, the existence of interaction effects among the mentioned risk factors may not be a significant concern. In other words, the primary effects of each variable may be more important, as these variables appear to play a role in CVD independently of each other.

Conflict of interests

The authors declare that they have no competing interests.

REFERENCES

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. Journal of the American College of Cardiology. 2020;76(25):2982-3021.

2. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, et al. Heart disease and stroke statistics—2023 update: a report from the American Heart Association. Circulation. 2023;147(8):e93-e621.

3. Sadeghi M, Haghdoost AA, Bahrampour A, Dehghani M. Modeling the burden of cardiovascular diseases in Iran from 2005 to 2025: the impact of demographic changes. Iranian journal of public health. 2017;46(4):506.

4. Aminorroaya A, Fattahi N, Azadnajafabad S, Mohammadi E, Jamshidi K, Rouhifard Khalilabad M, et al. Burden of non-communicable diseases in Iran: past, present, and future. Journal of Diabetes & Metabolic Disorders. 2020:1-7.

5. Kuulasmaa K, Tunstall-Pedoe H, Dobson A, Fortmann S, Sans S, Tolonen H, et al. Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations. The lancet. 2000;355(9205):675-87.

6. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. Journal of Computational and graphical Statistics. 2003;12(3):475-511.

7. Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B. Nonlinear estimation and classification: Springer Science & Business Media; 2013.

8. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. Genetic epidemiology. 2001;21(S1):S626-S31.

9. Cecil RLF, Goldman L, Schafer AI. Goldman's Cecil Medicine, Expert Consult Premium Edition--Enhanced Online Features and Print, Single Volume, 24: Goldman's Cecil Medicine: Elsevier Health Sciences; 2012.

10. Macdonald G. Harrison's Internal Medicine, -by AS Fauci, DL Kasper, DL Longo, E. Braunwald, SL Hauser, JL Jameson and J. Loscalzo. Wiley Online Library; 2008.

11. Collaboration ERF. Lipoprotein (a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. JAMA: the journal of the American Medical Association. 2009;302(4):412.

12. Collaboration ERF. C-reactive protein, fibrinogen, and cardiovascular disease prediction. New England Journal of Medicine. 2012;367(14):1310-20.

13. Humphrey LL, Fu R, Rogers K, Freeman M, Helfand M, editors. Homocysteine level and coronary heart disease incidence: a systematic review and meta-analysis. Mayo Clinic Proceedings; 2008: Elsevier.

14. Crawford MH, Education M-H. Current diagnosis & treatment in cardiology: McGraw Hill Medical; 2009.

15. 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(4):e0174944.

16. Hongzong S, Tao W, Xiaojun Y, Huanxiang L, Zhide H, Mancang L, et al. Support Vector Mechines Classification for Discriminating Coronary Heart Disease Patients from Non-coronary Heart Disease. West Indian Medical Journal. 2007;56(5):451.

17. Garcia M, Mulvagh SL, Bairey Merz CN, Buring JE, Manson JE. Cardiovascular disease in women: clinical perspectives. Circulation research. 2016;118(8):1273-93.

18. Keto J, Ventola H, Jokelainen J, Linden K, Keinänen-Kiukaanniemi S, Timonen M, et al. Cardiovascular disease risk factors in relation to smoking behaviour and history: a population-based cohort study. Open Heart. 2016;3(2):e000358.

19. England N, Ipsos M. GP patient survey. NHS England. 2015.

20. Hudson K, Lifton R, Patrick-Lake B, Burchard EG, Coles T, Collins R, et al. The precision medicine initiative cohort program—Building a Research Foundation for 21st Century Medicine. Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, ed. 2015.