Cell J
Cell Journal(Yakhteh)
Homepage: https://www.celljournal.org/
Original Article
Vol 25, No 8, August 2023, Pages: 536-545

Regularized Machine Learning Models for Prediction of Metabolic Syndrome Using *GCKR, APOA5,* and *BUD13* Gene Variants: Tehran Cardiometabolic Genetic Study

Nadia Alipour, Ph.D.[1], Anoshirvan Kazemnejad, Ph.D.[1]*, Mahdi Akbarzadeh, Ph.D.[2], Farzad Eskandari, Ph.D.[3], Asiyeh Sadat Zahedi, MSc.[2], Maryam S Daneshpour, Ph.D.[2]*

1. Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
2. Cellular and Molecular Endocrine Research Centre, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
3. Department of Statistics, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran

Abstract

Objective: Metabolic syndrome (MetS) is a complex multifactorial disorder that considerably burdens healthcare systems. We aim to classify MetS using regularized machine learning models in the presence of the risk variants of *GCKR, BUD13* and *APOA5*, and environmental risk factors.

Materials and Methods: A cohort study was conducted on 2,346 cases and 2,203 controls from eligible Tehran Cardiometabolic Genetic Study (TCGS) participants whose data were collected from 1999 to 2017. We used different regularization approaches [least absolute shrinkage and selection operator (LASSO), ridge regression (RR), elastic-net (ENET), adaptive LASSO (aLASSO), and adaptive ENET (aENET)] and a classical logistic regression (LR) model to classify MetS and select influential variables that predict MetS. Demographics, clinical features, and common polymorphisms in the *GCKR, BUD13,* and *APOA5* genes of eligible participants were assessed to classify TCGS participant status in MetS development. The models' performance was evaluated by 10-repeated 10-fold cross-validation. Various assessment measures of sensitivity, specificity, classification accuracy, and area under the receiver operating characteristic curve (AUC-ROC) and AUC-precision-recall (AUC-PR) curves were used to compare the models.

Results: During the follow-up period, 50.38% of participants developed MetS. The groups were not similar in terms of baseline characteristics and risk variants. MetS was significantly associated with age, gender, schooling years, body mass index (BMI), and alternate alleles in all the risk variants, as indicated by LR. A comparison of accuracy, AUC-ROC, and AUC-PR metrics indicated that the regularization models outperformed LR. Regularized machine learning models provided comparable classification performances, whereas the aLASSO model was more parsimonious and selected fewer predictors.

Conclusion: Regularized machine learning models provided more accurate and parsimonious MetS classifying models. These high-performing diagnostic models can lay the foundation for clinical decision support tools that use genetic and demographical variables to locate individuals at high risk for MetS.

Keywords: Classification, LASSO, Machine Learning, Metabolic Syndrome, Penalized Regression

Citation: **Alipour N, Kazemnejad A, Akbarzadeh M, Eskandari F, Zahedi AS, Daneshpour MS. Regularized machine learning models for prediction of metabolic syndrome using GCKR, APOA5, and BUD13 gene variants: Tehran cardiometabolic genetic study. Cell J. 2023; 25(8): 536-545. doi: 10.22074/CELLJ.2023.2000864.1294**

Introduction

Metabolic syndrome (MetS) is defined as a cluster of interrelated risk factors that directly increases the risk of cardiovascular diseases (CVD), type 2 diabetes, and other diseases (1). The prevalence of MetS is high in the USA, Europe, and Asian countries, including Iran (2, 3). In Iran, this prevalence is estimated to be 23.8% for those 20 or older and 10.98% for those under 20 years old. Thus, it imposes a considerable burden on the Iranian population and health system (4).

Appropriate predictive and classification models can help guide the interventions that aim to battle these conditions and their consequent complications and reduce their burden. In applying diagnostic models for MetS, two main aspects of prediction accuracy should be taken into consideration: i. The selection

Received: 25/April/2023, Revised: 08/June/2023, Accepted: 19/June/2023
***Corresponding Addresses: P.O.Box: 14115-111, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran**
P.O.Box: 19195-4763, Cellular and Molecular Endocrine Research Centre, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
Emails: kazem_an@modares.ac.ir, daneshpour@sbmu.ac.ir

Royan Institute
Cell Journal (Yakhteh)

of appropriate features and ii. Choice of classification algorithms (5). Different demographic and clinical characteristics can be used in MetS classification models. Family and twin research has shown that the components of MetS are inherited, and often occur together; this suggests a substantial genetic influence for the development of MetS. Heritability estimates for each MetS component are greater than 50%. Therefore, genetics play a significant role in predisposing individuals to its development. One potential application of this understanding is the use of genetic risk predictors to enhance the performance of diagnostic tools for multifactorial disorders (6).

The *GCKR* gene encodes glucokinase regulatory protein, an inhibitor of the glucose-metabolizing enzyme glucokinase. Glucokinase is responsible for regulating the uptake and storage of dietary glucose. A functional change in the glucokinase regulatory protein may considerably influence glucose metabolism (7). It has been demonstrated that rs1260326 is a functional variant that encodes a glucokinase regulatory protein (P446L-GKRP) which binds weakly to glucokinase at low glucose levels and indirectly results in increased glycolytic enzyme glucokinase (GCK) activity; also rs780094 and rs780093 are in linkage disequilibrium with this functional variant in different populations, including participants of the Tehran Cardiometabolic Genetic Study (TCGS) (8, 9). Two variants (rs780094-T and rs1260326-T) of *GCKR* are associated with enhanced glycolysis and lipogenesis (10). The association between the T allele of these common *GCKR* gene variants and an increased risk of MetS could be due to the relationship between these variants and elevated triglyceride (TG) levels, as a component of MetS (9).

Elevated TG levels also play a key role in the relationship between MetS and the *APOA5* and *BUD13* variants (11). Lipoprotein lipase activator, APOA5 protein, reduces very low-density lipoprotein synthesis and increases the absorption of lipoprotein remnants and insulin secretion in the liver. APOA5 is also associated with lower levels of free fatty acids and increased TG production (12). The *BUD13* and *APOA5* genes encode proteins in the APOA5 protein pathway. It has been hypothesized that the APOA5 protein's function can be altered by interactions with BUD13 and ZPR1 variants, which lead to elevated TG levels (13). Investigations in various populations, including European and Asian populations, have confirmed a correlation between the *APOA5* and *BUD13* variants and TG levels (14). These studies mainly focused on investigating the relationship between single nucleotide polymorphisms (SNPs) and MetS. Adding these variants to MetS classification models can boost their performance.

In addition, a proper classification model should be chosen. A traditional statistical model, like logistic regression (LR), is frequently used for classification tasks. However, LR is neither applicable nor suitable for classification tasks on correlated features and high-dimensional datasets. Recently, there has been an increasing interest in the use of regularized machine learning approaches to overcome these limitations (15) and also for triglycerides as local interactions within the 11q23.3 region (replicated significantly in NFBC1966. The most common regularization methods for feature selection tasks are the least absolute shrinkage and selection operator (LASSO), ridge regression (RR), elastic-net (ENET), adaptive LASSO (aLASSO), and adaptive ENET (aENET).

Therefore, we aim to compare regularized machine learning methods in feature selection and accurately classify MetS using demographic and clinical features of participants of the TCGS and their status in terms of selected variants of the *GCKR*, *BUD13*, and *APOA5* genes.

Materials and Methods

Design, setting, and participants

Participants for the current cohort study were selected from individuals who participated in the TCGS. The TCGS is a large population-based cohort study that has examined participants approximately every three years since 1999 in a family-based longitudinal framework. For this purpose, 15,005 participants of the first phase have been followed for more than 20 years in an attempt to monitor risk factors associated with non-communicable diseases to provide personalized medicine and the opportunity to present a patient-specific prevention plan before the onset of clinical symptoms. Details of this study protocol are available elsewhere (16). Of 15,005 individuals recruited to participate in TLGS from 1999 and who were followed up to 2017, 14,875 were selected to be part of TCGS (17).

For this study, individuals over 18 years of age who were not diagnosed with MetS during the first phase of the cohort were recruited. All participants with at least one follow-up measurement were included. Ultimately, 4,546 people (2,346 cases and 2,203 controls) were deemed eligible to be included in our analysis. Figure 1 presents a detailed flowchart of the patient recruitment.

The Ethics Committee of Tarbiat Modares University, Tehran, Iran approved this study (IR.MODARES.REC.1399.153). The Ethics Committee at Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran approved the design and conduct for all stages of the TCGS project. In each phase, all participants signed written informed consents.

Clinical and laboratory measurements

Demographic and clinical data that included age, gender, schooling years, physical activity, smoking status, marital status, and medication usage were collected

using certified questionnaires. Trained personnel took anthropometric measurements including weight, height, hip circumference, and waist circumference (WC). Details regarding anthropometric measurements and the collection of venous blood samples are available elsewhere (16). Enzymatic calorimetry (Pars Azmoun, Iran) was used for fasting blood sugar (FBS), TG, total blood cholesterol, and high-density lipoprotein cholesterol (HDL-C) levels. The variation (CV) coefficient range for TG, TC, FBS, and HDL-C was calculated at less than 5%. Friedwald's equation was used to calculate ow-density lipoprotein (LDL) cholesterol levels.

Outcome definition

Healthy (control) and unhealthy (case) participants were defined according to the definition provided by the Joint Interim Statement (JIS) criteria (18). According to the JIS committee, individuals have MetS if they had three or more of the following risk factors: i. Hypertension defined as a diastolic blood pressure ≥85 mmHg and systolic blood pressure ≥130 mmHg or use of antihypertensive medications, ii. Low HDL-C levels <40 mg/dL for males and <50 mg/dL for females while fasting or use of lipid-lowering medications, iii. High levels of fasting serum TG (≥150 mg/dL) or use of medications for TG, iv. High levels of fasting plasma glucose (≥100 mg/dL) or use of diabetes medications. v. Central obesity, which was defined as a WC of ≥90 cm for both genders, according to the Iranian National Committee of Obesity guidelines (19). Individuals who were not diagnosed with MetS during any of the six phases of TCGS comprised the control group.

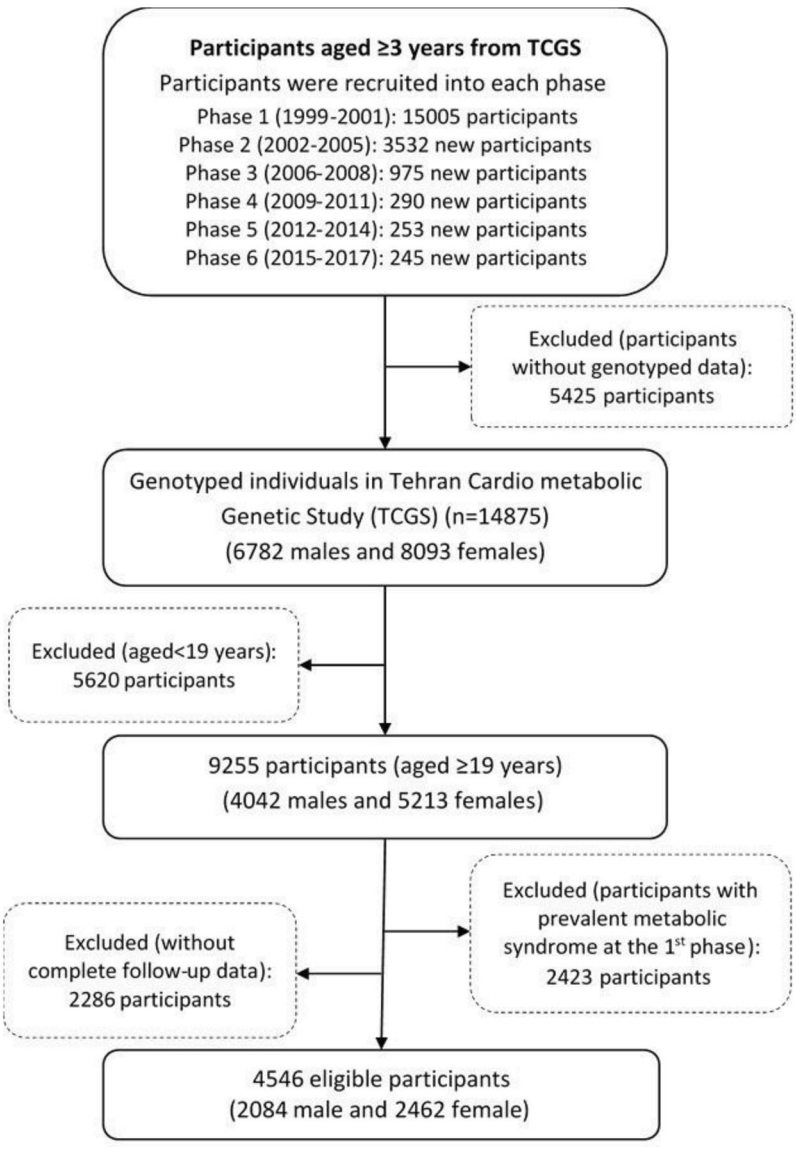

Fig.1: Flowchart of the study selection process. The term "new participants" refers to individuals who enrolled in the study at different phases of this study and were not present during the earlier phases. TCGS; Tehran Cardiometabolic Genetic Study.

Genotyping and single nucleotide polymorphisms selection

The standard proteinase K/salting-out approach was used to extract genomic DNA samples from the venous blood samples' buffy-coat. The quality and quantity of each DNA sample were evaluated using a Thermo Scientific NanoDrop 1000 Spectrophotometer. Samples with low quality and concentration (DNA purification in the range of 1.7<A260/A280<2) were omitted. DNA samples of TCGS participants were genotyped at the deCODE Genetics, Inc. (Reykjavik, Iceland) with HumanOmniExpress-24-v1-0 bead chips that contained 649 932 SNP loci with an average mean distance of 4 kb, and according to the manufacturer's specifications (Illumina, Inc., San Diego, CA, USA). Allele frequency, divergence from Hardy-Weinberg equilibrium (HWE) in the control group, and individual-level missingness were verified using PLINK (version 1.9) (20) before the association analysis was performed. Seven SNPs located on the *GCKR* (rs780094, rs1260326, and 780093), *BUD13* (rs12292921, rs180326), and *APOA5* (rs651821, rs2266788) genes were selected based on a demonstrated strong association with MetS in the TCGS population (9, 11).

Statistical analysis

Descriptive data statistics were expressed as mean ± standard deviation (SD) for continuous variables and frequencies (percent) for categorical variables. Differences between the MetS and non-MetS groups by the study covariates and genotypes for both responders and non-responders by sex were calculated using the Student's t test or chi-square test.

The additive LR model was used to assess the association between MetS with candidate SNPs (selected variants of the *GCKR*, *BUD13*, and *APOA5*) considering the covariates of age, gender, body mass index (BMI), schooling years, physical activity, smoking status, and marital status. In the additive model, the BB, Bb, and bb genotypes were recoded to 0, 1, and 2, respectively. For feature selection and MetS classification, LR was performed through the forward stepwise method.

We applied regularized LR (RLR) to select relevant features for MetS classification and prediction models. RLR models avoid the overfitting problem by penalizing the model complexity and adding a nonnegative regularization term to the log-likelihood function, consequently, shrinking the values of regression coefficients (21, 22). This study used a 10-fold cross-validation method to select the optimal λ value for our models. We took into consideration several regularization terms that have been previously proposed (21-25) and applied five popular regularized machine learning models to select relevant features for MetS prediction models, including the LASSO (21), RR (23), ENET (25), aLASSO (24), and aENET. A detailed explanation of the regularization methods and their features was added in the supplementary material and Table S1 (See Supplementary Online Information at www.celljournal.org).

Evaluation method

A 10-repeated 10-fold cross-validation method was used to evaluate the performance of the models. This method splits the data into ten randomly selected subsets. Each subset of 10% (testing set) is used to assess the model exclusively trained on the remaining 90% of individuals (training set). The data sample is shuffled in each repetition, which results in a different set of sample data after each split. The performance metrics gathered from ten repeated cross-validations are then averaged to calculate the overall model performance metrics. The main advantage of applying this procedure in assessing model performance is that it reduces the performance estimates variance (26). The proportion of cases and controls in each subset was maintained equal to the primary sample proportion, so each subset's status could represent the underlying population's status.

Performance metrics of sensitivity, specificity, classification accuracy, and the area under the receiver operating characteristic curve (AUC-ROC) are shown in Figure S1 (See Supplementary Online Information at www.celljournal.org) and the area under the precision-recall curve (AUC-PR) in Figure S2 (See Supplementary Online Information at www.celljournal.org). Finally, the Delong test was used to compare the AUC-ROC values of the different regression models (27). We used R version 4.0.2 for implementing the mentioned methods, along with the packages "*glmnet*", "*caret*", "*msaenet*", "*pROC*", and "*PRROC*" (28-32). All the analyses were considered at a 0.05 level of significance.

Results

Study population characteristics

Out of the total 4546 participants in the study, 54.16% were females. A total of 2343 (50.38%) of the TCGS adult participants had MetS. The mean BMI for patients with MetS was 27.07 ± 4.15 kg/m^2, which indicated that MetS patients were generally overweight. The majority of participants (53.32%) had never smoked. Table 1 shows the baseline features of participants

The genotype distribution of all seven SNPs in the control subjects was in HWE (P>0.05). Table 2 displays the baseline characteristics and common SNPs of the *GCKR*, *BUD13*, and *APOA5* genotypes for both responders and non-responders by gender. Based on the results, the number of women in both responders and non-responders was more than men. There were no significant differences (P>0.05) between the TCGS responders and non-responders in both males and females (P>0.05) unless patterns of smoking (P<0.05), marital status (P<0.05), and *BUD13* and *APOA5* variant distributions were observed in both women and men.

Table 1: Comparison of baseline characteristics and common SNPs of the *GCKR, BUD13*, and *APOA5* genotypes of MetS in healthy control and unhealthy groups

Variables	Unhealthy (MetS)	Healthy (Non-MetS)	P value	Test statistic
Group size	2343 (50.38)	2203 (49.62)		
Age (Y)	40.33 ± 12.92	33.16 ± 12.47	<0.001	19.01‡
Schooling years	9.20 ± 4.35	10.44 ± 4.63	<0.001	9.310‡
BMI (kg/m^2)	27.07 ± 4.15	23.82 ± 3.94	<0.001	27.04‡
Physical activity	587 ± 370	459 ± 115	<0.001	15.54‡
Sex			<0.001	112.28§
Male	1252 (53.4)	832 (37.8)		
Female	1091 (46.6)	1371 (62.2)		
Smoking status			<0.001	39.55§
Never smoked	1175 (50.15)	1249 (56.7)		
Former smoker	151 (6.44)	73 (3.31)		
Current smoker	345 (14.72)	258 (11.71)		
Second-hand	672 (28.68)	623 (28.28)		
Marital status			<0.001	148.6§
Divorced	24 (1)	19 (0.9)		
Married	1946 (83.1)	1549 (70.3)		
Single	314 (13.4)	609 (27.6)		
Widowed	59 (2.5)	26 (1.2)		
GCKR				
rs1260326			0.035	6.675§
CC	662 (28.3)	689 (31.3)		
TC	1134 (48.4)	1054 (47.8)		
TT	547 (23.3)	460 (20.9)		
rs780094			0.028	7.157§
CC	672 (28.7)	706 (32)		
TC	1138 (48.6)	1045 (47.4)		
TT	533 (22.7)	452 (20.5)		
rs780093			0.005	10.46§
CC	666 (28.4)	693 (31.5)		
TC	1125 (48)	1072 (48.7)		
TT	552 (23.6)	438 (19.9)		
BUD13				
rs12292921			0.128	4.058§
GG	20 (0.9)	12 (0.5)		
GT	311 (13.3)	259 (11.8)		
TT	2012 (85.9)	1932 (87.7)		
rs180326			0.016	8.324§
GG	451 (19.2)	408 (18.5)		
GT	1211 (51.7)	1068 (48.5)		
TT	681 (29.1)	727 (33)		
APOA5				
rs651821			0.006	10.25§
TT	1626 (69.4)	1614 (73.3)		
CT	649 (27.7)	546 (24.8)		
CC	68 (2.9)	43 (2)		
rs2266788			0.007	9.722§
GG	54 (2.3)	32 (1.5)		
GA	584 (24.9)	490 (22.2)		
AA	1705 (72.8)	1681 (76.3)		

Significant differences were observed in SNP information of *GCKR*, *BUD13*, and *APOA5* genotypes, and independent variables between healthy and unhealthy participants. Data are presented as mean ± SD or n (%). ‡; Student's t test, §; Chi-square test, SNP; Single nucleotide polymorphisms, MetS; Metabolic syndrome, BMI; Body mass index, and SD; Standard deviation.

Table 2: Comparison of baseline characteristics and *GCKR, BUD13, APOA5* genotype of study participants and non-responders

Variables	Male				Female			
	Responders	Non-responders	P value	Test statistic	Responders	Non-responders	P value	Test statistic
Group size	2084 (71.51)	830 (28.48)			2462 (67.17)	1203 (32.82)		
Age (Y)	39.44 ± 14.59	41.48 ± 15.31	<0.001	3.358‡	34.67 ± 11.45	37.06 ± 14.6	<0.001	5.404‡
Schooling years	10.03 ± 4.66	9.91 ± 4.41	0.582	0.636‡	9.60 ± 4.41	9.32 ± 4.36	0.0701	1.811‡
BMI (Kg/m^2)	24.87 ± 3.88	25.12 ± 4.38	0.025	1.511‡	26.02 ± 4.68	26.63 ± 4.92	<0.001	3.643‡
Physical activity	622 ± 108	524.49 ± 101.92	0.05	22.34‡	443 ± 293	384.48 ± 182.76	<0.001	6.349‡
Smoking status			0.152	5.283§			0.0646	7.238§
Never smoked	775 (37.19)	283 (38.34)			1649 (66.98)	717 (59.60)		
Former smoker	209 (10.03)	88 (11.92)			15 (0.61)	13 (1.08)		
Current smoker	537(25.77)	196 (26.56)			66 (2.68)	35 (2.91)		
Second-hand	563(27.02)	171 (23.17)			732 (29.73)	371 (30.84)		
Marital status			0.200	4.637§			<0.001	28.62§
Divorced	10 (0.5)	1(0.05)			33 (1.3)	18 (1.50)		
Married	1579 (75.8)	649(29.99)			1916 (77.8)	954 (79.30)		
Single	491 (23.6)	176 (8.13)			432 (17.5)	155 (12.88)		
Widowed	4 (0.2)	3 (0.14)			81 (3.3)	76 (6.32)		
GCKR								
rs1260326			0.218	3.04§			0.581	1.083§
CC	634 (30.4)	265 (31.93)			717 (29.1)	370 (30.76)		
TC	987 (46.9)	402 (48.43)			1210 (49.1)	574 (47.71)		
TT	472 (22.6)	163 (49.64)			535 (21.7)	259 (21.53)		
rs780094			0.290	2.472§			0.959	0.083§
CC	643 (30.9)	251 (30.24)			735 (29.9)	358 (29.76)		
TC	980 (47)	414 (49.88)			1203 (48.9)	584 (48.55)		
TT	461 (22.1)	165 (19.88)			524 (21.3)	261 (21.70)		
rs780093			0.192	3.297§			0.053	5.855§
CC	643 (30.9)	179 (21.57)			716 (29.1)	280 (23.28)		
TC	975 (46.8)	315 (37.95)			1222 (49.6)	404 (33.58)		
TT	466 (22.4)	125 (15.06)			524 (21.3)	214 (17.79)		
BUD13								
rs12292921			0.369	0.831§			<0.001	34.24§
GG	19 (0.9)	9 (1.08)			13 (0.5)	8 (0.67)		
GT	257 (12.3)	107 (12.89)			313 (12.7)	241 (20.03)		
TT	1808 (86.8)	714 (86.02)			2136 (86.8)	954 (79.30)		
rs180326			<0.001	31.88§			<0.001	30.75§
GG	405 (19.4)	108 (13.01)			454 (18.4)	154 (12.80)		
GT	1055 (50.6)	512 (61.69)			1224 (49.7)	706 (58.69)		
TT	624 (29.9)	210 (25.30)			784 (31.8)	343 (28.51)		
APOA5								
rs651821			<0.001	22.06§			0.799	0.448§
TT	1490 (71.5)	663 (79.88)			1750 (71.1)	863 (71.74)		
CT	541 (26)	149 (17.95)			654 (26.6)	309 (25.69)		
CC	53 (2.5)	18 (2.17)			58 (2.4)	31 (2.58)		
rs2266788			0.004	11.09§			<0.001	2407.93§
GG	40 (1.9)	21 (2.53)			46 (1.9)	863 (71.74)		
GA	482 (23.1)	237 (28.55)			592 (24)	309 (25.69)		
AA	1562 (75)	572 (68.92)			1824 (74.1)	31 (2.58)		

There were no significant differences between responders and non-responders in males and females other than higher BMI in male non-responders and different smoking and marital status distribution between female responders and non-responders. Significant differences between distribution of the *BUD13* and *APOA5* variants were observed between responders and non-responders in both genders. Data are presented as mean ± SD or n (%). ‡; Student's t test, §; chi-square test, SD; Standard deviation, BMI; Body mass index. Student's t test for quantitative variables and chi-square test for categorical variables were applied.

Feature selection

Table 3 lists the subset of features selected for each of the six assessed models: LASSO, ridge, ENET, aLASSO, aENET, and LR. aLASSO produced the most parsimonious model with six features: BMI, marital status, and gene variants of *GCKR* (rs780093), *BUD13* (rs12292921), as well as gene variants of *APOA5* (rs651821 and rs2266788). Features whose coefficients have a larger absolute value have a more significant effect on predicting MetS. Likewise, lower values show less influence on prediction. BMI had a large effect, despite being selected by all models. LR analysis revealed that males were 0.14 times less at risk for development of MetS than females. Increased number of school attendance years had a significant inverse relationship with the odds of developing MetS (OR=0.02). In contrast, variants of the *BUD13* (rs12292921, rs180326) and *GCKR* (rs780094, rs780093) genes, gender, and BMI strongly affected MetS. Analysis of *GCKR* polymorphisms (rs780094, rs1260326, and rs780093) demonstrated a significant association of MetS with rs780093. This relationship can be attributed to the higher frequency of minor T alleles in patients affected by MetS. Similarly, *BUD13* polymorphisms, specifically rs12292921 and rs180326, exhibited a significant association with MetS. Patients with MetS have a greater prevalence of minor G alleles, which is the leading cause of this relationship. Analysis of *APOA5* polymorphisms revealed an association of MetS with rs651821 and rs2266788. The presence of minor C alleles of rs651821 and minor G alleles rs2266788 were found to be significantly more frequent in MetS patients than in healthy individuals.

Predictive performance

Figure 2 provides an overview of the classification performance achieved by regularized and classic LR techniques. The evaluation is based on the mean value obtained from conducting a 10-repeat 10-fold cross-validation method. The results show that regularization methods outperformed the LR model based on several metrics (Fig.2).

Overall, the aENET showed higher classification accuracy (mean=0.748) and sensitivity (mean=0.763), which was very similar to aLASSO. Figures S1 and S2 (See Supplementary Online Information at www.celljournal.org) show the AUC-ROC and AUC-PR curves of the models. The classification accuracy is the most exact and simplest measure of model performance and it showed the highest values for the aLASSO and aENET models. AUC-ROC and AUC-PR summarize the trade-off between sensitivity and specificity or precision and recall at different probability thresholds, respectively. Based on these indices the aENET model is the exact model for the classification of individuals. aLASSO gave parsimony results.

We used the Delong et al. (27) method to compare the six calculated AUCs. The results of the pairwise comparison are shown in Table 4. All five regularization machine learning models outperformed LR, and we confirmed there were no differences between the regularization models. In contrast, we observed significant differences between the AUCs of the regularized and classic LR models ($P<0.05$).

Table 3: Feature selection using different regularized machine learning and classical LR models

Variable	LR	LASSO	Ridge	ENET	aLASSO	aENET
Age (Y)	0.025	0.025	0.024	0.025		0.023
Gender	0.864	0.834	0.774	0.855		0.928
BMI (kg/m^2)	0.222	0.168	0.188	0.210	0.013	0.214
Schooling years	-0.021	-0.017	-0.018	-0.017		-0.019
Physical activity			0.0009			
Marital status			0.113	0.060	0.057	0.058
Smoking status		-0.013	0.019			
GCKR						
rs1260326	0.036		0.016			0.077
rs780094	-1.20		-0.062			-0.095
rs780093	0.093	0.012	0.182	0.141	-0.076	0.169
BUD13						
rs12292921	-0.321	-0.210	-0.257	-0.224	-0.207	-0.309
rs180326	0.235	0.035	0.067	0.037		0.061
APOA5						
rs651821	0.055	0.109	0.111	0.043	0.139	0.169
rs2266788	-0.051	-0.172	-0.224	-0.275	-0.053	-0.194

The coefficient of significant variables is based on the LR model and selected features using the regularized machine learning models. aLASSO was the parsimony model, among others. LR; Logistic regression, LASSO; Least absolute shrinkage and selection operator, ENET; Elastic-net, aENET; Adaptive ENET, aLASSO; Adaptive LASSO, and BMI; Body mass index.

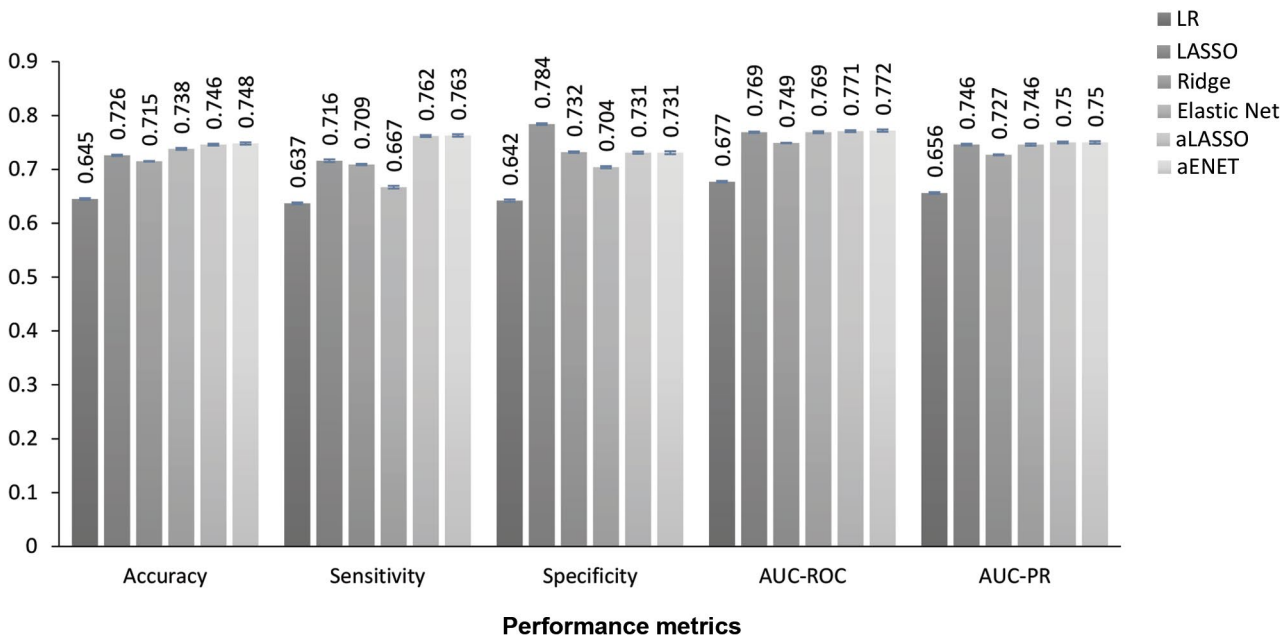

Fig.2: Performance metrics of the regularized machine learning models. AUC-ROC; Area under the receiver operating characteristic curve, AUC-PR; Area under the precision-recall curve, LR; Logistic regression, LASSO; Least absolute shrinkage and selection operator, Elastic-net; ENET, aLASSO; Adaptive LASSO, and aENET; Adaptive ENET, Regularized machine learning models outperform the LR model.

Table 4: Pairwise comparison of receiver operating characteristic curves (ROC)

Models	D statistics	Significance level (P value)	Models	D statistics	Significance level (P value)
LR vs. LASSO	7.89	<0.05	LASSO vs. aENET	0.215	0.64
LR vs. Ridge	6.59	<0.05	Ridge vs. ENET	0.168	0.79
LR vs. ENET	7.76	<0.05	Ridge vs. aLASSO	0.117	0.67
LR vs. aLASSO	7.52	<0.05	Ridge vs. aENET	0.129	0.71
LR vs. aENET	7.67	<0.05	ENET vs. aLASSO	0.143	0.67
LASSO vs. Ridge	0.215	0.61	ENET vs. aENET	0.214	0.85
LASSO vs. ENET	0.131	0.89	aLASSO vs. aENET	0.210	0.63
LASSO vs. aLASSO	0.152	0.76			

Using the Delong et al. (27) method, we confirmed differences between LR and the regularized machine learning models. However, there were no discriminative differences between the regularization models. LR; Logistic regression, LASSO; Least absolute shrinkage and selection operator, aLASSO; Adaptive LASSO, ENET; Elastic-net, and aENET; Adaptive ENET.

Discussion

The main objective of this study was to classify MetS using various penalized machine learning models and LR in the presence of risk variants in the *GCKR*, *BUD13* and *APOA5* genes, and determine the crucial variables for the development of MetS. Regularized machine learning models proved parsimonious MetS predicting models and outperformed LR while selecting fewer features. Our result confirmed the superiority of regularized machine learning models over LR; these findings supported the results of a study that compared the models to characterise vitamin D deficiency in a hypertensive obese population (33). Our results were further verified by Kim et al. (34), who reported that LASSO LR outperformed stepwise LR to predict breast malignancies.

For all three models, the final selected variables were BMI, rs2266788, rs651821, rs12292921, and rs780093. The findings of previous studies that employed data mining and machine learning techniques to examine the association between MetSd and demographic and lifestyle factors have identified BMI as one of the most crucial predictors of MetS. This underscores the

importance of maintaining a healthy weight in preventing the development of MetS (35, 36).

All of the models in our study selected four functional variants of the *GCKR* (rs780093), *BUD13* (rs12292921), and *APOA5* (rs2266788, rs651821) genes as predictors of MetS. This result is consistent with previous studies that investigated the relationship between *GCKR*, *BUD13*, and *APOA5* polymorphisms and MetS in different populations (11, 12, 35, 37). The results of the present study support our previous findings when we evaluated machine learning models to predict MetS in the presence of *GCKR* risk variants (35). Some of the regularized methods selected three other SNPs. However, not chosen as a significant predictor of an outcome does not necessarily negate the biological or clinical importance of a specific exposure variable or its causal significance (38).

aENET and aLASSO provided higher performance measures, but this difference was not significant compared to other regularized machine learning models. On the other hand, the aLASSO model was the most parsimonious, with the least number of selected features. This finding can have substantial clinical importance. High-performing models built on fewer variables are easier to interpret and more practical because of the typical time constraints that impede gathering patient data in most clinical settings. Due to their comparable performance with different feature sets, each of these five regularized machine learning models can have practical advantages. With only six predictors, the aLASSO is the most advantageous.

The current study has several advantages compared to other studies to classify MetS. First, we used our models on a relatively large sample that represented its underlying population. Our feature set comprised genetic and environmental risk factors that properly captured the multifactorial nature of MetS and maintained clinical intelligibility and practicality throughout the research design and implementation. More complicated modelling techniques were avoided and not considered to be a shortcoming of this effort. Although other classification methods, particularly those based on machine learning algorithms like support vector machines or random forest, might perform better than regularized machine learning methods on the same data sets (35, 39), it is important to remember that these high-performing methods might not be easily implemented in many clinical settings.

In terms of clinical issues, a high-precision model with fewer variables to predict individuals' status is valuable because of the time constraints with obtaining patient information. The utility of the regularized machine learning models was demonstrated based on the results and the models' performance parameter estimation. It is important to acknowledge that the findings of this study may not be applicable to other populations due to potential differences in demographics, cultural norms, and environmental factors. Therefore, caution should be exercised when attempting to generalize these results to other populations.

Conclusion

In order to maximize the potential impact of public health interventions in reducing the burden of prevalent diseases like MetS, it is critical to focus resources on people who are at a higher risk for developing these diseases or who already are afflicted. Conventional statistical models tend to be unreliable when predicting multifactorial disorders that have numerous potential independent environmental and genetic risk factors. In contrast, modern machine learning algorithms such as penalized regressions can significantly improve predictive accuracy in clinical matters compared to conventional models. In this study, we compared prediction models for MetS by using demographic, lifestyle, and genetic data (risk variants of *GCKR*, *BUD13*, and *APOA5* genes) from TCGS patients. Our findings indicate that penalized machine learning models, specifically aENET and aLASSO, can provide highly effective MetS prediction models. These models can play a crucial role in preventing future CVD, cancers, or other related complications if they are integrated with decision support tools or used in future research.

This work was the first step to apply the risk score as a modern method for disease prediction. The key focus in the TCGS is to identify the best prediction model(s) for various diseases, particularly MetS, which is a complex disorder caused by multiple factors. In order to achieve this goal, we tested penalized machine learning techniques and compared them to LR model on known genes within our database in an attempt to compare their predictive abilities.

Acknowledgments

This paper was a collaborative project between the Faculty of Medical Sciences at Tarbiat Modares University and the Research Institute of Endocrine Sciences at Shahid Beheshti University of Medical Sciences. The authors gratefully acknowledge Tarbiat Modares University Faculty of Medical Sciences for their financial support (grant number: Med85794) and the staff and participants of the TCGS for providing data. We especially express our appreciation to deCODE Genetics, Inc. (Reykjavik, Iceland) for their scientific support. The authors declare no conflict of interest.

Authors' Contributions

N.A., M.S.D., A.S.Z.; Designed the study. A.S.Z.; Data curation. N.A., M.A.; Programming and software. N.A.; Formal analysis, and wrote the original draft. A.K., F.E.; Validation. A.K., M.S.D.; Wrote, reviewed and edited the manuscript. A.K., M.S.D.; Provided supervision. All authors read and approved the final manuscript.

References

1. Kassi E, Pervanidou P, Kaltsas G, Chrousos G. Metabolic syn-

drome: definitions and controversies. BMC Med. 2011; 9: 48.
2. Jahangiry L, Khosravi-Far L, Sarbakhsh P, Kousha A, Entezarmahdi R, Ponnet K. Prevalence of metabolic syndrome and its determinants among Iranian adults: evidence of IraPEN survey on a biethnic population. Sci Rep. 2019; 9(1): 7937.
3. Hirode G, Wong RJ. Trends in the prevalence of metabolic syndrome in the United States, 2011-2016. JAMA. 2020; 323(24): 2526-2528.
4. Mazloomzadeh S, Rashidi Khazaghi Z, Mousavinasab N. The prevalence of metabolic syndrome in iran: a systematic review and meta-analysis. Iran J Public Health. 2018; 47(4): 473-480.
5. Lee YH, Bang H, Kim DJ. How to establish clinical prediction models. Endocrinol Metab (Seoul). 2016; 31(1): 38-44.
6. Abou Ziki MD, Mani A. Metabolic syndrome: genetic insights into disease pathogenesis. Curr Opin Lipidol. 2016; 27(2): 162-171.
7. Matschinsky FM. Banting Lecture 1995. A lesson in metabolic regulation inspired by the glucokinase glucose sensor paradigm. Diabetes. 1996; 45(2): 223-241.
8. Yuan F, Gu Z, Bi Y, Yuan R, Niu W, Ren D, et al. The association between rs1260326 with the risk of NAFLD and the mediation effect of triglyceride on NAFLD in the elderly Chinese Han population. Aging (Albany NY). 2022; 14(6): 2736-2747.
9. Zahedi AS, Akbarzadeh M, Sedaghati-Khayat B, Seyedhamzehzadeh A, Daneshpour MS. GCKR common functional polymorphisms are associated with metabolic syndrome and its components: a 10-year retrospective cohort study in Iranian adults. Diabetol Metab Syndr. 2021; 13: 20.
10. Fernandes Silva L, Vangipurapu J, Kuulasmaa T, Laakso M. An intronic variant in the GCKR gene is associated with multiple lipids. Sci Rep. 2019; 9(1): 10240.
11. Masjoudi S, Sedaghati-Khayat B, Givi NJ, Bonab LNH, Azizi F, Daneshpour MS. Kernel machine SNP set analysis finds the association of BUD13, ZPR1, and APOA5 variants with metabolic syndrome in Tehran Cardio-metabolic Genetics Study. Sci Rep. 2021; 11(1): 10305.
12. Oh SW, Lee JE, Shin E, Kwon H, Choe EK, Choi SY, et al. Genome-wide association study of metabolic syndrome in Korean populations. PLoS One. 2020; 15(1): e0227357.
13. Kim HK, Anwar MA, Choi S. Association of BUD13-ZNF259-APOA5-APOA1-SIK3 cluster polymorphism in 11q23.3 and structure of APOA5 with increased plasma triglyceride levels in a Korean population. Sci Rep. 2019; 9(1): 8296.
14. Aung LH, Yin RX, Wu JZ, Wu DF, Wang W, Li H. Association between the MLX interacting protein-like, BUD13 homolog and zinc finger protein 259 gene polymorphisms and serum lipid levels. Sci Rep. 2014; 4: 5565.
15. Wei W, Gyenesei A, Semple CA, Haley CS. Properties of local interactions and their potential value in complementing genome-wide association studies. PLoS One. 2013; 8(8): e71203.
16. Daneshpour MS, Fallah MS, Sedaghati-Khayat B, Guity K, Khalili D, Hedayati M, et al. Rationale and design of a genetic study on cardiometabolic risk factors: protocol for the tehran cardiometabolic genetic study (TCGS). JMIR Res Protoc. 2017; 6(2): e28.
17. Daneshpour MS, Hedayati M, Sedaghati-Khayat B, Guity K, Zarkesh M, Akbarzadeh M, et al. Genetic identification for non-communicable disease: findings from 20 years of the Tehran lipid and glucose study. Int J Endocrinol Metab. 2018; 16 Suppl 4: e84744.
18. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, et al. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation. 2009; 120(16): 1640-1645.
19. Azizi F, Khalili D, Aghajani H, Esteghamati A, Hosseinpanah F, Delavari A, et al. Appropriate waist circumference cut-off points among Iranian adults: the first report of the Iranian national committee of obesity. Arch Iran Med. 2010; 13(3): 243-244.
20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3): 559-575.
21. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996; 58(1): 267-288.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005; 67(2): 301-320.
23. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1): 55-67.
24. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006; 101(476): 1418-1429.
25. Zou H, Zhang HH. ON The adaptive elastic-net with a diverging number of parameters. Ann Stat. 2009; 37(4): 1733-1751.
26. Kuhn M, Johnson K. Over-fitting and model tuning. In: Kuhn M, Johnson K, editors. Applied predictive modeling. 1st ed. New York: Springer; 2013.
27. 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(3): 837-845.
28. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; 33(1): 1-22.
29. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008; 28(5): 1-26.
30. Xiao N, Xu QS. Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection. J Stat Comput Simul. 2015; 85(18): 3755-3765.
31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12: 77.
32. Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics. 2015; 31(15): 2595-2597.
33. Garcia-Carretero R, Vigil-Medina L, Barquero-Perez O, Mora-Jimenez I, Soguero-Ruiz C, Goya-Esteban R, et al. Logistic LASSO and Elastic net to characterize vitamin D deficiency in a hypertensive obese population. Metab Syndr Relat Disord. 2020; 18(2): 79-85.
34. Kim SM, Kim Y, Jeong K, Jeong H, Kim J. Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. Ultrasonography. 2018; 37(1): 36-42.
35. Akbarzadeh M, Alipour N, Moheimani H, Zahedi AS, Hosseini-Esfahani F, Lanjanian H, et al. Evaluating machine learning-powered classification algorithms which utilize variants in the GCKR gene to predict metabolic syndrome: Tehran cardio-metabolic genetics study. J Transl Med. 2022; 20(1): 164.
36. Huang YC. The application of data mining to explore association rules between metabolic syndrome and lifestyles. Health Inf Manag. 2013; 42(3): 29-36.
37. Jasim AA, Al-Bustan SA, Al-Kandari W, Al-Serri A, AlAskar H. Sequence analysis of APOA5 among the Kuwaiti population identifies association of rs2072560, rs2266788, and rs662799 with TG and VLDL levels. Front Genet. 2018; 9: 112.
38. Schooling CM, Jones HE. Clarifying questions about "risk factors": predictors versus explanation. Emerg Themes Epidemiol. 2018; 15: 10.
39. Guo S, Lucas RM, Ponsonby AL; Ausimmune Investigator Group. A novel approach for prediction of vitamin d status using support vector regression. PLoS One. 2013; 8(11): e79970.