antioxidants

Article

A Machine Learning Algorithm for Quantitatively Diagnosing Oxidative Stress Risks in Healthy Adult Individuals Based on Health Space Methodology: A Proof-of-Concept Study Using Korean Cross-Sectional Cohort Data

Youjin Kim [1,†], Yunsoo Kim [1,†], Jiyoung Hwang [2], Tim J. van den Broek [3], Bumjo Oh [4], Ji Yeon Kim [5], Suzan Wopereis [3], Jildau Bouwman [3,*] and Oran Kwon [1,2,*]

[1] Department of Nutritional Science and Food Management, Ewha Womans University, 52 Ewhayeodae-gil, Seodeamun-gu, Seoul 03760, Korea; Youjin.Kim631782@tufts.edu (Y.K.); sookim726@gmail.com (Y.K.)

[2] Department of Nutritional Science and Food Management, Graduate Program in System Health Science and Engineering, Ewha Womans University, 52 Ewhayeodae-gil, Seodeamun-gu, Seoul 03760, Korea; cindy.jyhwang@gmail.com

[3] Netherlands Organization for Applied Scientific Research (TNO), Department of Microbiology and Systems Biology, Utrechtseweg 48, 3704 HE Zeist, The Netherlands; tim.vandenbroek@tno.nl (T.J.v.d.B.); suzan.wopereis@tno.nl (S.W.)

[4] Boramae Medical Center, Department of Family Medicine, Seoul Metropolitan Government-Seoul National University, 20 Boramae-ro 5-gil, Dongjak-gu, Seoul 07061, Korea; bo39@snu.ac.kr

[5] Department of Food Science and Technology, Seoul National University of Science and Technology, 232 Gongneung-ro, Nowon-gu, Seoul 01811, Korea; jiyeonk@seoultech.ac.kr

[*] Correspondence: jildau.bouwman@tno.nl (J.B.); orank@ewha.ac.kr (O.K.); Tel.: +31-88-866-1678 (J.B.); +82-2-3277-6860 (O.K.)

[†] These authors contributed equally to this work.

Citation: Kim, Y.; Kim, Y.; Hwang, J.; van den Broek, T.J.; Oh, B.; Kim, J.Y.; Wopereis, S.; Bouwman, J.; Kwon, O. A Machine Learning Algorithm for Quantitatively Diagnosing Oxidative Stress Risks in Healthy Adult Individuals Based on Health Space Methodology: A Proof-of-Concept Study Using Korean Cross-Sectional Cohort Data. *Antioxidants* **2021**, *10*, 1132. https://doi.org/10.3390/antiox10071132

Academic Editors: Claudia Fiorillo and Matteo Becatti

Received: 26 May 2021
Accepted: 14 July 2021
Published: 16 July 2021

Abstract: Oxidative stress aggravates the progression of lifestyle-related chronic diseases. However, knowledge and practices that enable quantifying oxidative stress are still lacking. Here, we performed a proof-of-concept study to predict the oxidative stress status in a healthy population using retrospective cohort data from Boramae medical center in Korea (n = 1328). To obtain binary performance measures, we selected healthy controls versus oxidative disease cases based on the "health space" statistical methodology. We then developed a machine learning algorithm for discrimination of oxidative stress status using least absolute shrinkage and selection operator (LASSO)/elastic net regression with 10-fold cross-validation. A proposed fine-tune model included 16 features out of the full spectrum of diverse and complex data. The predictive performance was externally evaluated by generating receiver operating characteristic curves with area under the curve of 0.949 (CI 0.925 to 0.974), sensitivity of 0.923 (CI 0.879 to 0.967), and specificity of 0.855 (CI 0.795 to 0.915). Moreover, the discrimination power was confirmed by applying the proposed diagnostic model to the full dataset consisting of subjects with various degrees of oxidative stress. The results provide a feasible approach for stratifying the oxidative stress risks in the healthy population and selecting appropriate strategies for individual subjects toward implementing data-driven precision nutrition.

Keywords: elastic net regularized generalized linear model; diagnostic model; oxidative stress; composite biomarker

1. Introduction

In general, aging comprises interconnected physiological processes and simultaneous unfavorable body composition and metabolic dysfunction [1]. These changes may cause persistent oxidative stress, leading to a state of chronic low-grade inflammation [2]. If not adequately controlled, oxidative stress and chronic low-grade inflammation may result in structural and functional abnormalities, leading to diet- and lifestyle-related chronic diseases, such as metabolic syndrome, cardiovascular disease, neurodegenerative disease, and certain cancers [1–3]. Therefore, diagnosing and identifying oxidative stress risks at an early stage is essential to optimize health and wellbeing and alleviate the increasing burden of diet- and lifestyle-related chronic diseases [4].

Many studies have identified variables, including cigarette smoking [5], aging [3], lipid peroxidation product malondialdehyde (MDA) [6,7], and body mass index (BMI) [8,9], as associated factors with oxidative stress. However, in the diagnostic setting, the response cannot be based on a single variable due to the multifactorial components found within the person [10]. According to Califf (2018), a composite biomarker may help understand the subtle changes that describe the processes and their interactions instead. Moreover, each of the multiple variables may play a critical role in the summative outcome of oxidative stress, allowing monitoring of disease progression at an early stage [11]. Researchers have used many statistical or graphical techniques like pattern recognition, principal component analysis, and partial least squares discrimination analysis to analyze and visualize multiple variables at a time [12]. However, these techniques often define the axis based on the variation or line that best separates the defined groups, thus having no biological meaning [13]. Bouwman et al. (2012) recently proposed a statistical visualization method, named 'health space', that addresses this issue by projecting individual subjects' health status based on predefined biological processes determined by multivariate parameterization.

The rapid development in big data analysis and artificial intelligence has recently begun to unlock clinically relevant information from cohort data, journals, and clinical practices hidden in a large volume of these, assisting clinical practice by providing up-to-date medical information or reducing diagnostic and therapeutic errors [14–16]. However, to the best of our knowledge, no studies have applied machine learning (ML) algorithms to serve as a starting point to reduce the risk of oxidative stress-related chronic diseases in advance and answer concerning data-driven precision nutrition. We tackle this problem by exploring a proof-of-concept study to test the utilization of the health space methodology for extensive population studies. Furthermore, we used ML approaches to develop a model that discriminates oxidative stress risks with strong prediction power and interpretability by selecting valuable features from the full spectrum of diverse and complex data. Subsequently, we validated the proposed diagnostic model in a separate hold-out test set of subjects. This study may initiate and facilitate evidence-based data-driven precision nutrition.

2. Materials and Methods

2.1. Study Population

This study used a retrospective cohort design to analyze 2454 subjects, who first came to Seoul Metropolitan Government–Seoul National University Boramae Medical Center (Seoul, Republic of Korea) for regular general health check-ups between 1 April 2015 and 31 August 2018. A total of 1328 subjects were included after excluding (1) subjects with missing data on any independent variables selected for analysis (n = 618), (2) subjects aged under 20 years old (n = 5), (3) same persons visited twice (n = 117), and (4) patients diagnosed with diseases excluding oxidative diseases (n = 386). The Institutional Review Boards of Seoul National University Boramae Medical Center (approval number 20140929/26-2014-118/102) and Ewha Womans University (approval number 86-8) approved this study. We received written informed consent from all subjects and performed all procedures following the relevant guidelines and regulations.

2.2. Data Collection

We extracted a total of 43 features from the record and developed a structured database. Data on general characteristics included age, sex, smoking status, alcohol consumption, and all diagnosed diseases. The overall diet quality and physical activity were determined by the recommended food score (RFS) [17] and validated Korean version of the International Physical Activity Questionnaire [18], respectively. Anthropometric data included body fat percentage and BMI. Biochemical analyses of serum samples included albumin, alkaline phosphatase (ALP), blood bilirubin, blood urea nitrogen (BUN), creatinine, C-reactive protein (CRP), γ-glutamyl transferase (GGT), glutamate oxaloacetate transaminase (GOT), glutamate pyruvate transaminase (GPT), glycosylated hemoglobin (HbA1c), low-density lipoprotein cholesterol (LDL-C), total protein, total cholesterol (TC), and uric acid (UA). Complete blood count (CBC) data included basophils, eosinophils, erythrocyte sedimentation rate (ESR), hematocrit (Hct), hemoglobin (Hb), lymphocytes, mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), monocytes, neutrophils, platelet count, platelet distribution width (PDW), red blood cell count (RBC), red blood cell distribution width (RDW), and white blood cell count (WBC).

2.3. MDA Analysis

We analyzed the MDA levels in plasma, erythrocyte, and urine using an HPLC system (Shiseido, Tokyo, Japan) equipped with a fluorescence detector (emission = 527 nm, excitation = 551 nm). A Capcell Pak C18 UG120 column (4.6 mm × 250 mm, 5 µm particle size; Shiseido) was used for separation under isocratic elution with 50 mM phosphate buffer [19–22].

2.4. Development and Validation of ML Algorithm

A schematic of the analysis pipeline is presented in Figure 1. We first split the eligible dataset (n = 1328) into training (n = 911) and hold-out (n = 417) datasets after removing the duplicate samples. The training dataset consisted of subjects who received a regular general health check-up at the Seoul Metropolitan Government–Seoul National University Boramae Medical Center in 2015–2016. The most recent dataset of the 2017–2018 period was used as the hold-out test set. Notably, the separate and independent hold-out dataset was used only for validating the final model to ensure no bias in the model development.

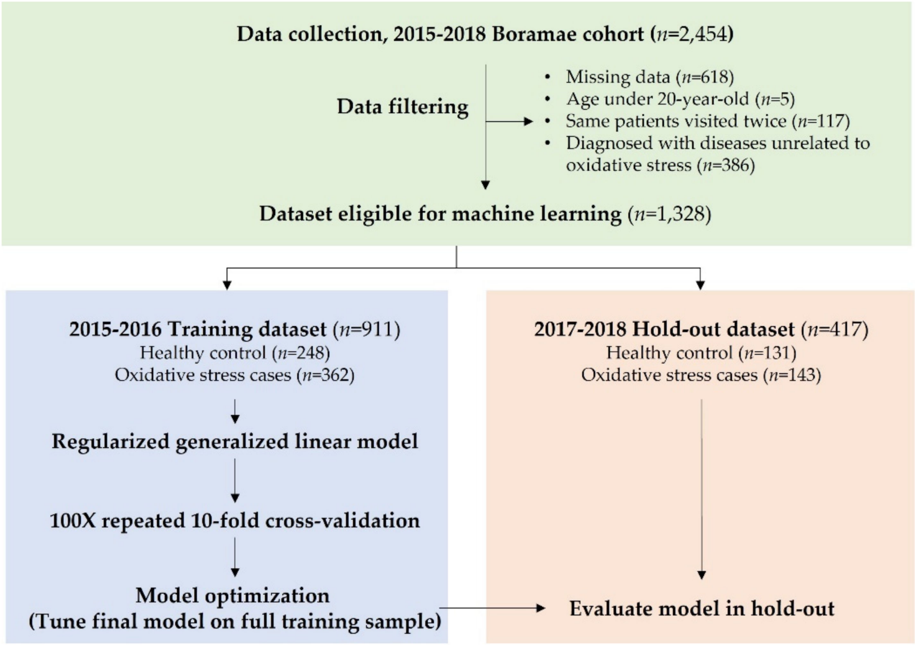

Data collection, 2015-2018 Boramae cohort (n=2,454)
Data filtering
Missing data (n=618)
Age under 20-year-old (n=5)
Same patients visited twice (n=117)
Diagnosed with diseases unrelated to oxidative stress (n=386)
Dataset eligible for machine learning (n=1,328)
2015-2016 Training dataset (n=911)
Healthy control (n=248)
Oxidative stress cases (n=362)
Regularized generalized linear model
100X repeated 10-fold cross-validation
Model optimization
(Tune final model on full training sample)
2017-2018 Hold-out dataset (n=417)
Healthy control (n=131)
Oxidative stress cases (n=143)
Evaluate model in hold-out

Figure 1. Schematic of the analysis pipeline. The eligible datasets were split into training and hold-out validation datasets. The least absolute shrinkage and selection operator (LASSO)/elastic net regression algorithm was trained using the training dataset and further evaluated in the hold-out validation dataset.

To obtain binary performance measures, we formed two reference groups in the training and hold-out datasets by adopting a health space model [13]. The healthy controls were defined as subjects having neither metabolic syndrome nor its components nor physician-diagnosed medical conditions. Extensive phenotyping of oxidative stress-related diseases was made based on the previously published data. As a result, the oxidative disease cases were defined as subjects with metabolic syndrome [1], dyslipidemia [9], hypertension [9], intermediate coronary syndrome [5], stroke [7], diabetes mellitus [23], or liver cirrhosis [24], among all subjects diagnosed as having diseases (Supplementary Materials Figure S1). We used the disease terminology presented by the Human Phenotype Ontology [25].

Considering the high-dimensional and collinear characteristics of the data, we employed the regularized generalized linear model (GLM) with the least absolute shrinkage or selection operator (LASSO) and elastic net penalties using the R package "glmnet" [26]. The mixing parameter alpha was set to 0.5, reflecting the balance between the ridge and LASSO penalties of the two approaches [27]. The resulting multivariate regression model was validated by 10-fold stratified cross-validation (CV) to avoid overfitting [28]. This procedure was repeated 100 times to improve stability by increasing the number of evaluations from 10 to 1000 [29]. Then, based on the one-standard-error rule, we calculated the optimal value of lambda (λ) and the minimum misclassification rate (lambda.1se) to tune the elastic net and LASSO regressions [30,31]. The best-performing model was identified by comparing the two best candidates derived from each elastic net and LASSO regression using the area under the receiver operating characteristic curve (AUC). Finally, external validation of the best performing model was carried out by calculating AUC, specificity, sensitivity, accuracy, negative predictive value (NPV), positive predictive value (PPV), positive clinical utility index (CUI+), and negative clinical utility index (CUI-) [32,33].

Furthermore, we investigated calibration to confirm whether the predicted probabilities agree with the observed probabilities [34]. To this end, the final prediction algorithm was applied to the entire dataset (n = 1328), stratified into four categories according to the number of metabolic syndrome risk factors (0, 1, and 2) and the presence of oxidative stress diseases. Next, we applied our final prediction algorithm to the subjects excluded from this study due to other conditions unrelated to oxidative stress. Then, Duncan's post hoc test was used to delineate group differences further.

2.5. Statistics

All statistical analyses were performed using the R software (version. 3.6.1; R Foundation for Statistical Computing) [35]. The results were expressed descriptively as means and standard deviation (SD) for continuous variables and number (percentage) for categorical variables. Differences between healthy controls and oxidative disease cases were compared by the Student's t-test for continuous variables and Chi-square test for categorical variables. Statistical significance was set at $p < 0.05$.

3. Results

3.1. Characteristics of the Reference Groups

Out of 1328 subjects, 884 samples were extracted as healthy controls (n = 379, 28.5%) and oxidative disease cases (n = 505, 38%) to develop and validate the ML algorithm discriminating oxidative stress risks in the healthy population. Table 1 compares the 43 features derived from these two reference groups in the training and hold-out dataset, respectively. The results indicate the separable features of the groups. In both datasets, the subjects in the oxidative disease cases were older, with a higher percentage of male subjects and smokers than the healthy controls. The oxidative disease cases had significantly higher BMI and body fat percentages than the healthy controls. This trend was similar for the ALP, BUN, creatinine, CRP, GGT, GPT, HbA1c, and UA levels. In contrast, albumin, bilirubin, LDL-C, TC, and total protein levels did not differ between the two groups. For CBC data, Hb, Hct, MCHC, RBC, and WBC were significantly higher in the oxidative disease cases than in the healthy controls.

Table 1. Comparison between healthy controls and oxidative disease cases included in the training and hold-out datasets for the machine learning algorithm discriminating oxidative stress risks in the healthy population.

Features	2015–2016 Training Set (*n* = 610)		*p*-Value	2017–2018 Hold-Out Set (*n* = 274)		*p*-Value
	Healthy Controls (*n* = 248)	Oxidative Disease Cases (*n* = 362)		Healthy Controls (*n* = 131)	Oxidative Disease Cases (*n* = 143)	
General characteristics (10)						
Age (years)	42.7 ± 10.4	52.5 ± 10.2	<0.0001	38.3 ± 10.0	52.4 ± 10.1	<0.0001
Sex (males; *n*, %)	101 (40.7)	264 (72.9)	<0.0001	47 (35.9)	102 (71.3)	<0.0001
Current smoker (*n*, %)	23 (9.3)	97 (26.8)	<0.0001	5 (3.8)	15 (10.5)	0.034
Smoking duration (y)	4.5 ± 9.0	11.7 ± 13.1	<0.0001	3.1 ± 6.8	9.3 ± 12.6	<0.0001
Smoking pack-year	3.3 ± 7.6	9.5 ± 12.7	<0.0001	2.4 ± 6.4	9.3 ± 15.7	<0.0001
Current drinker (*n*, %)	145 (58.5)	224 (61.9)	0.397	78 (59.5)	82 (57.3)	0.712
RFS	21.6 ± 8.3	24.7 ± 8.7	<0.0001	20.5 ± 8.7	22.5 ± 9.0	0.069
Physical activity (m/d)	128.5 ± 120.1	133.9 ± 118.9	0.581	119.2 ± 113.8	81.0 ± 107.0	0.005
BMI (kg/m^2)	21.5 ± 2.2	25.6 ± 4.1	<0.0001	21.2 ± 2.2	25.4 ± 3.7	<0.0001
Body fat (%)	25.5 (6.4)	28.6 (6.9)	<0.0001	26.0 (6.4)	29.3 (6.5)	<0.0001
Biochemical characteristics (14)						
Albumin (g/dL)	4.3 ± 0.2	4.4 ± 0.2	0.089	4.4 ± 0.2	4.5 ± 0.2	0.079
AP (IU/L)	62.5 ± 17.8	72.6 ± 18.5	<0.0001	62.3 ± 17.2	70.1 ± 17.8	<0.0001
Bilirubin (mg/dL)	1.2 ± 0.5	1.1 ± 0.4	0.129	1.1 ± 0.4	1.1 ± 0.4	0.237
BUN (mg/dL)	12.8 ± 3.5	13.9 ± 3.4	<0.0001	11.6 ± 2.8	13.8 ± 3.3	<0.0001
Creatinine (µmol/L)	68.2 ± 14.1	77.0 ± 15.5	<0.0001	66.1 ± 12.6	77.6 ± 15.8	<0.0001
CRP (mg/dL)	0.1 ± 0.2	0.2 ± 0.3	<0.0001	0.1 ± 0.3	0.2 ± 0.4	<0.0001
GT (IU/L)	18.4 ± 15.0	38.3 ± 38.2	<0.0001	17.7 ± 13.7	39.7 ± 36.0	<0.0001
GOT (IU/L)	22.6 ± 6.8	30.5 ± 18.4	<0.0001	26.7 ± 56.0	31.2 ± 17.0	0.372
GPT (IU/L)	18.9 ± 12.0	35.5 ± 32.4	<0.0001	18.6 ± 12.3	34.2 ± 24.2	<0.0001
Glycosylated Hb (%)	5.4 ± 0.3	6.0 ± 0.9	<0.0001	5.3 ± 0.3	5.9 ± 0.9	<0.0001
LDL-C (mmol/L)	3.0 ± 0.7	3.1 ± 0.9	0.223	3.0 ± 0.7	3.2 ± 1.1	0.028
TC (mmol/L)	5.0 ± 0.8	5.1 ± 1.1	0.293	5.0 ± 0.8	5.2 ± 1.1	0.089
Total protein (g/dL)	7.1 ± 0.4	7.2 ± 0.4	0.283	7.2 ± 0.3	7.3 ± 0.4	0.081
Uric acid (mg/dL)	4.8 ± 1.2	5.6 ± 1.4	<0.0001	4.7 ± 1.1	5.6 ± 1.3	<0.0001
Complete blood count data (16)						
Basophils (%)	0.4 ± 0.3	0.4 ± 0.3	0.504	0.5 ± 0.3	0.4 ± 0.3	0.273
Eosinophils (%)	2.6 ± 2.2	2.8 ± 2.2	0.273	2.7 ± 2.0	2.9 ± 2.5	0.419
ESR (mm/h)	9.0 ± 8.2	9.6 ± 8.4	0.380	9.4 ± 11.9	11.8 ± 11.1	0.090
Hematocrit (%)	41.4 ± 4.2	44.2 ± 3.8	<0.0001	42.7 ± 3.7	44.9 ± 3.7	<0.0001
Hemoglobin (g/dL)	13.9 ± 1.6	15.0 ± 1.5	<0.0001	14.2 ± 1.5	15.1 ± 1.4	<0.0001
Lymphocytes (%)	35.3 ± 8.8	35.4 ± 8.1	0.883	36.4 ± 7.9	36.3 ± 7.6	0.893
MCH (pg)	30.3 ± 1.9	30.8 ± 1.5	<0.0001	30.0 ± 1.9	30.4 ± 1.3	0.094
MCHC (g/dL)	33.5 ± 1.0	34.0 ± 0.9	<0.0001	33.2 ± 0.9	33.5 ± 0.9	0.002
MCV (fL)	90.4 ± 4.4	90.8 ± 3.9	0.209	90.6 ± 4.7	90.7 ± 3.4	0.830
Monocytes (%)	5.7 ± 1.5	5.6 ± 1.4	0.424	5.8 ± 1.4	5.9 ± 1.5	0.709
Neutrophils (%)	56.0 ± 9.7	55.8 ± 8.6	0.775	54.7 ± 8.4	54.6 ± 8.5	0.905
Platelets ($\times 10^3$/µL)	249.1 ± 51.9	247.9 ± 53.4	0.781	258.5 ± 54.0	258.4 ± 86.8	0.995
PDW (%)	11.8 ± 1.6	11.9 ± 1.5	0.412	11.8 ± 1.5	12.2 ± 1.6	0.016
RBC count ($\times 10^6$/µL)	4.6 ± 0.4	4.9 ± 0.5	<0.0001	4.7 ± 0.4	5.0 ± 0.4	<0.0001
RDW (%)	13.1 ± 1.0	12.9 ± 0.6	0.014	13.1 ± 1.1	12.9 ± 0.6	0.188
WBC count ($\times 10^3$/µL)	5.2 ± 1.3	6.2 ± 1.7	<0.0001	5.2 ± 1.3	6.1 ± 1.8	<0.0001
High-performance liquid chromatography analysis data (3)						
Plasma MDA (mM)	4.4 ± 2.4	5.0 ± 2.5	0.009	4.3 ± 2.2	4.8 ± 2.0	0.083
Erythrocyte MDA (mM)	6.1 ± 6.2	6.3 ± 6.9	0.601	8.8 ± 8.6	9.4 ± 7.7	0.564
Urine MDA (mM)	2.9 ± 1.8	2.7 ± 1.5	0.380	2.7 ± 1.7	2.5 ± 1.3	0.207

Values are mean ± SD or number (percentage). Differences between the groups were compared using Student's *t*-test for continuous variables and Chi-square tests for categorical variables. AP, alkaline phosphatase; BMI, body mass index; BUN, blood urea nitrogen; CRP, C-reactive protein; GOT, glutamate oxaloacetate transaminase; GPT, glutamate pyruvate transaminase; GT, γ-Glutamyl transferase; LDL-C, low-density lipoprotein cholesterol; ESR, erythrocyte sedimentation rate; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; MDA, malondialdehyde; PDW, platelet distribution width; RDW, red blood cell distribution width; RFS, recommended food score; TC, total cholesterol.

3.2. Developing an ML Algorithm for Discriminating Oxidative Stress Risks in the Healthy Population

The heatmap presented in Figure 2 shows all features and corresponding coefficients obtained from the training dataset using logistic regression with elastic net (upper) and LASSO (lower) penalties. Ten-fold CV was run 100 times for each penalty. The color scale beneath the heatmap represents a range of coefficient values, where blue is negative and red is positive. Twenty-one out of a total of 43 features were extracted at least once within the 200 individual replications. Age, BMI, GGT, GPT, Hb, HbA1c, and WBC were the most consistently extracted features across all 200 models.

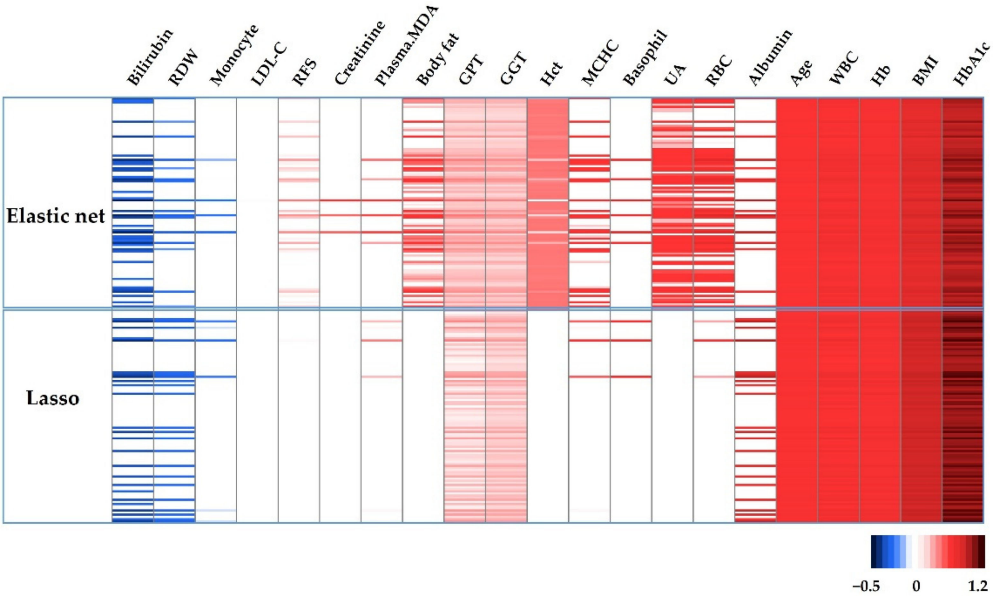


Figure 2. Regression coefficients for the features selected in 200 different 10-CV least absolute shrinkage and selection operator (LASSO)/elastic net regularized generalized linear models. The color scale beneath the heatmap represents a range of coefficient values from negative (blue) to red (positive). CV, cross-validation; RDW, red blood cell distribution width; LDL-C, low-density lipoprotein cholesterol; RFS, recommended food score; MDA, malondialdehyde; GPT, glutamate pyruvate transaminase; GGT, γ-glutamyl transferase; Hct, hematocrit; MCHC, mean corpuscular hemoglobin concentration; UA, uric acid; RBC, red blood cell count; WBC, white blood cell count; Hb, hemoglobin; BMI, body mass index; HbA1c, glycosylated hemoglobin.

The CV results are presented in Figure 3A, depicting the mean squared prediction error against log λ with one-standard error bars. The vertical dashed lines indicate the location of the minimum misclassification error (lambda.1se) selected by a "one-standard-error" rule. The elastic net regression gives lambda.1se = 0.0253, and the LASSO regression yields lambda.1se = 0.0167, indicating that LASSO produced a more regularized model compared with the elastic net model. Figure 3B compares the performance of the elastic net and LASSO models using AUC, validating the above findings that the LASSO model (AUC 0.949, 95% confidence interval [CI] 0.925–0.974) performed slightly better than the elastic net model (AUC 0.948, 95% CI 0.924–0.973). The best LASSO regression model contains 16 features (age, plasma MDA, BMI, RFS, HbA1c, GPT, GGT, bilirubin, albumin, WBC, RBC, Hb, RDW, monocytes, basophils, and MCHC). The feature with the highest negative coefficient value was bilirubin, while those with the highest positive coefficient value were HbA1c and albumin. In contrast, the best elastic net regression model contains four more features, including creatinine, UA, body fat percentage, and LDL-C.

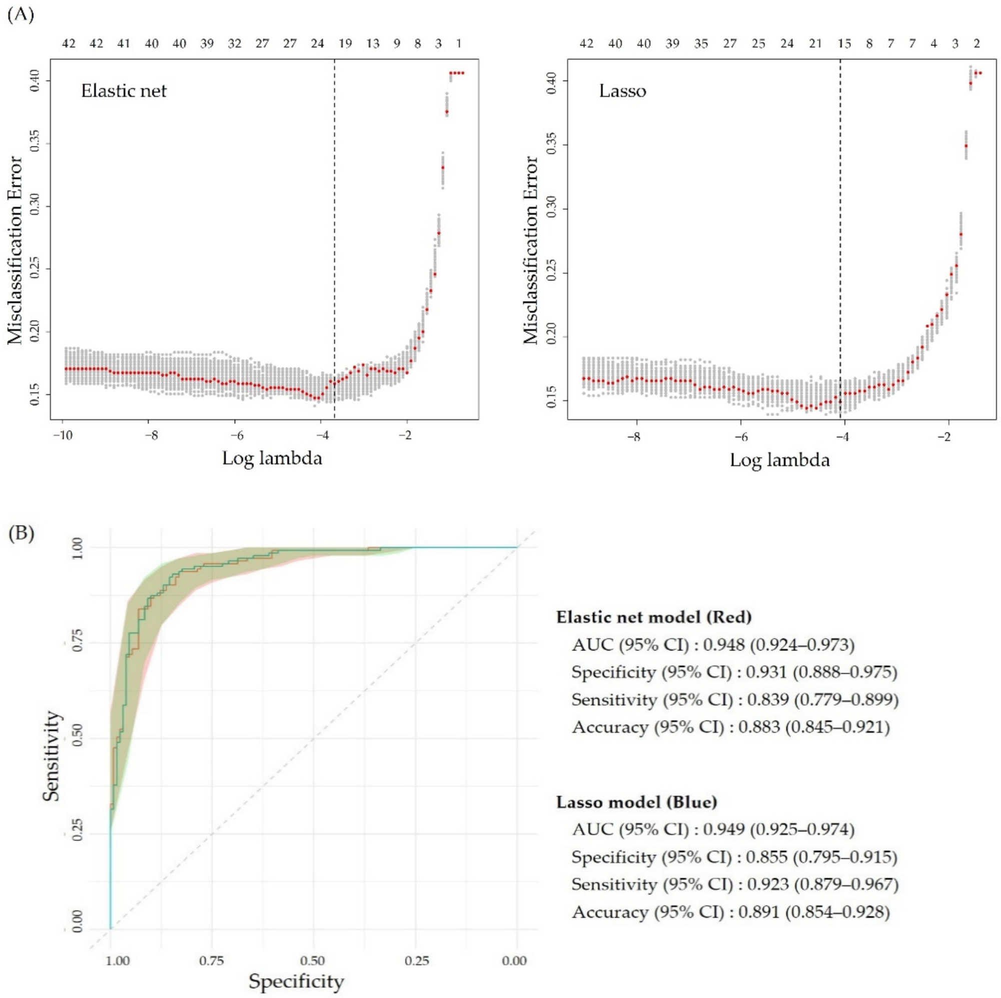


Figure 3. Selection of the best performing model. (**A**) Cross-validation estimate of the mean squared prediction error for elastic net (left) and least absolute shrinkage and selection operator (LASSO, right), as a function of log λ. Grey bars are one-standard errors, and the vertical dashed lines are the location of the minimum misclassification error. Numbers on the top of the panel indicate the number of features selected. (**B**) Receiver operating characteristic curve for the elastic net and LASSO model on hold-out samples. The diagonal line represents the reference line of 0.5. AUC, area under the curve; CI, confidence interval.

3.3. Internal and External Validation of the Best Performing Model

Figure 4A shows the diagnostic performance of our final best performing model in the training dataset (n = 610), which consisted of 248 healthy controls and 362 oxidative disease cases. The result was excellent, presenting an AUC of 0.935 (95% CI, 0.916–0.953), specificity of 0.839 (95% CI, 0.793–0.884), sensitivity of 0.881 (95% CI, 0.848–0.915), accuracy of 0.864 (95% CI, 0.837–0.891), NPV of 0.829 (95% CI, 0.782–0.875), PPV of 0.889 (95% CI, 0.856–0.921), CUI+ of 0.783 (95% CI, 0.743–0.823), and CUI- of 0.695 (95% CI, 0.655–0.735). Figure 4B demonstrates the outstanding discrimination performance of our final model validated using the hold-out dataset (n = 274), which comprised 131 healthy controls and 143 oxidative disease cases. The result was even better, with an AUC of 0.949 (95% CI, 0.925–0.974), specificity of 0.855 (95% CI, 0.795–0.915), sensitivity of 0.923 (95% CI, 0.879–0.967), accuracy of 0.891 (95% CI, 0.854–0.928), NPV of 0.911 (95% CI, 0.86–0.961), PPV of 0.874 (95% CI, 0.821–0.927), CUI+ of 0.807 (95% CI, 0.747–0.867), and CUI- of 0.779 (95% CI, 0.73–0.827). Statistically, the healthy controls in the training set and those in the hold-out set had compatible variables, except for age, albumin, BUN, HbA1c, total protein, Hct, MCHC, RBC, and erythrocyte MDA. However, when we compare those values with the reference guideline level, they were within the normal ranges (Table S1).

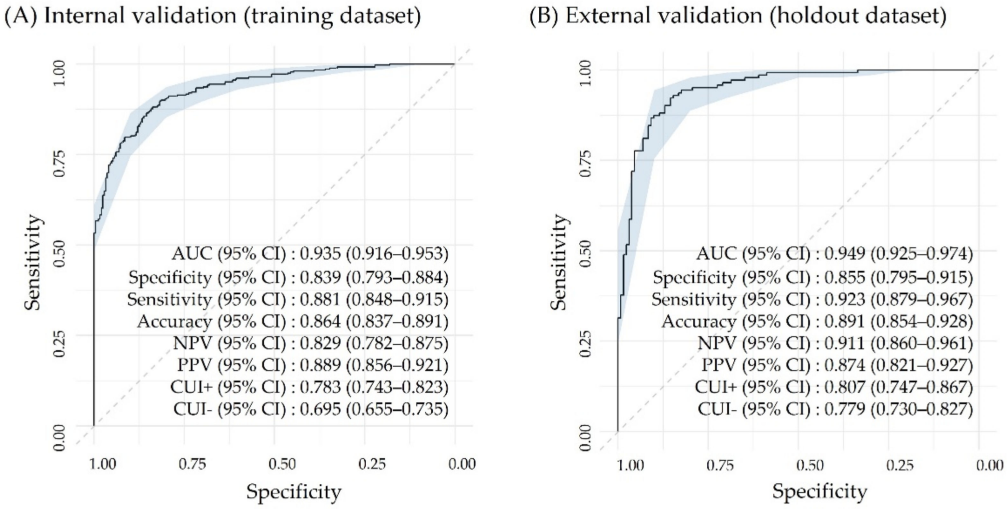


Figure 4. Performance of the best performing machine learning model on two datasets. (**A**) Training dataset and (**B**) hold-out dataset. The diagonal line represents the reference line of 0.5. Light blue indicates 95% CIs of the ROC curve estimate. CI, confidence interval; ROC, receiver operating characteristic.

3.4. Application of the Best Performing Model to All Subjects in the Whole Dataset for Testing Discrimination Power

Figure 5A is a violin plot illustrating the diagnostic values for oxidative stress of all individuals stratified by four categories based on the number of metabolic syndrome risk factors and the presence of oxidative stress diseases. The results confirmed that our final model can suitably define healthy and oxidative disease categories, which seems to be better than the traditional separation based on clinical diagnosis or the metabolic syndrome definition. Moreover, our final model can identify individuals with higher metabolic risks as having higher oxidative stress risks toward the oxidative disease cases (p for trend < 0.001). The result presented in Figure 5B shows the diagnostic values for oxidative stress of the subjects excluded from this study due to the presence of other diseases unrelated to oxidative stress. The other disease groups colored in purple were significantly different or at least tended to differ from either the healthy controls or the oxidative stress cases, but it was challenging to calibrate the apparent magnitude of oxidative stress risks.

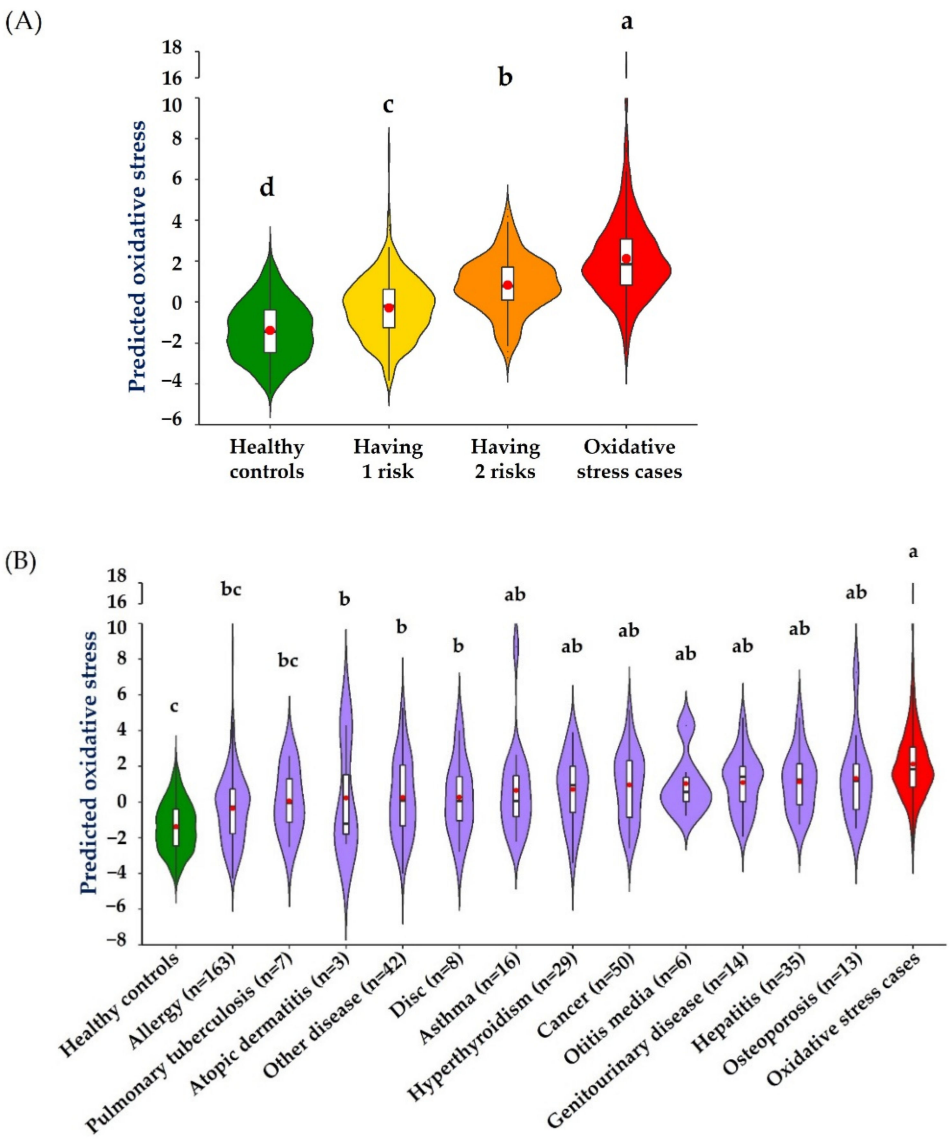


Figure 5. Application of the final model for qualitative diagnosis of oxidative stress risks. (**A**) Calibration of oxidative stress risks of subjects having zero (n = 379), one (n = 282), and two metabolic risk factors (n = 162) and oxidative disease cases (n = 505). (**B**) Application of oxidative stress risks of subjects with other diseases not related to oxidative stress (n = 386) in comparison to the healthy controls (n = 379) and oxidative disease cases (n = 505). The violin plot represents the distribution of data by using the kernel density function, and the width of the violin plot represents the sample size at this level. The red dot on each box plot represents the mean value. The box in the violin plot represents the median and quartile; the extension from the thin black line represents the 95% confidence interval. The different lowercase letters indicate a significant difference between the health status by Duncan's test ($p < 0.05$).

4. Discussion

Given the need to better discriminate high-risk individuals for diet- and lifestyle-related chronic diseases in the general population, we developed and validated an ML model to diagnose individuals' oxidative stress risks at an early stage. Many prediction models for disease diagnosis or prognosis have emerged [10,36], but they do not enable individuals to monitor disease deterioration. The present study was initiated as proof of the health space model, a statistical model for analyzing and visualizing the treatment effects of functional foods in individual subjects based on predefined biological processes [13]. The health space model was applied to depict subjects' oxidative stress status by evaluating the summative outcomes of biological processes, thus avoiding erroneous conclusions [11]. Furthermore, it was expanded to derive binary samples of healthy controls and oxidative disease cases from the Boramae cohort data.

As oxidative stress is complex and multifactorial, a composite biomarker might facilitate diagnosing and risk-stratifying subjects with high oxidative stress levels than any single biomarker [6,37–41]. The current study used ML models in the multivariate statistical analysis rather than the traditional statistical approaches. The advantages of ML models over traditional statistical approaches include their ability to consider interactions between features and explore combinations that might not be apparent [42]. Among many algorithms, the GLM-based technique has mainly been used for the text mining of electronic health reports and developing a prediction model in health care, providing a simple and interpretable description [43]. However, the performance of GLMs has been indicated as unsatisfactory because they are prone to overfitting and sensitive to outliers [44]. Alternatively, a graph learning-matching network (GLMNet) that fits a GLM via penalized maximum likelihood can be used to overcome the limitations of naïve GLMs [28,45,46]. In this current study, we applied the two representative regularization methods called LASSO and elastic net. Our results showed that the LASSO regression better predicted oxidative stress risks than the elastic net regression.

In a data mining algorithm, proper internal validation is essential to prevent model overfitting and reduce potential false-positive findings [47]. The split-sample approach is a straightforward and popular method in which the training data are randomly split into two parts for developing a model and measuring its performance. The CV is a more sophisticated approach with the advantage of intuitively being regarded as an extension of the split-sample method [48]. In the current study, we applied CV for internal validation with a 10% fraction to test a model developed on 90% of the sample in this context. This procedure was repeated 100 times in the training dataset to improve CV stability. Such stratification is ideal for keeping the test dataset intact and bringing it out only at the end of the data analysis. However, because the test dataset was used repeatedly in this study, we determined the regularization parameters to choose the most parsimonious model whose error is no more than one standard error [31,48]. As a result, we obtained a final transparent and easy-to-interpret diagnostic model composed of a combination of 16 features representing the composite of anthropometrical, biochemical, and clinical data.

We note that we reserved a separate and independent hold-out dataset for external validation of the final model. AUC analysis was performed to assess the final model performance, presenting an AUC value of 0.949 in the validation dataset versus 0.935 in the training dataset. In the context of discrimination, AUC values of 1.0, 0.9–0.99, 0.8–0.89, 0.7–0.79, 0.51–0.69, and 0.5 can be interpreted as perfect, excellent, good, fair, poor, and of no value, respectively [49]. Therefore, our model may provide appropriate selection criteria for individuals to implement data-driven precision nutrition. Contrary to our result, some studies that explored a global oxidative stress index (Oxidative-INDEX) did not reach beyond the good prediction level. Park et al. [50] created an oxidative stress score consisting of MDA, oxidized low-density lipoprotein, and 8-epi-prostaglandin F2α. The AUC value was 0.75 when combined with single nucleotide polymorphisms to predict obesity in the Korean population. In another study, the Oxidative-INDEX was calculated in patients with coronary artery disease by measuring overall pro- and antioxidant exposure balance after z-score standardization [51]. It showed significant associations of the score with diabetes, smoking habit, hypercholesterolemia, aging, and CRP at the multivariate regression analysis, suggesting the potential use of the Oxidative-INDEX in preventing, diagnosing, and treating coronary artery disease. However, validation remained undone, and further investigation was needed in the general population.

There is increasing interest in the application of ML for nutritional science to prevent and improve diet- and lifestyle-related chronic diseases. However, to our knowledge, no prior published studies included the application of ML algorithms for quantifying health conditions and providing appropriate strategies to individuals for implementing data-driven precision nutrition in the context of our study. Our model was able to quantitatively stratify oxidative stress burden, as expected. It could even distinguish healthy individuals from most individuals with diseases unrelated to oxidative stress, although to a relatively lesser degree of preciseness. However, this study had several limitations. First, even though the model performance was excellent, as indicated by CV discrimination and external validation, there was partial overlap between the groups categorized by the number of metabolic syndrome risk factors (0, 1, and 2) and the presence of oxidative stress diseases. Other statistical considerations may have to be enjoined to solve this problem. Second, this study was conducted using the dataset obtained from adults residing in the Republic of Korea. Thus, the results of this study may not be generalizable to all ethnic populations. Further investigations are underway to develop and validate an oxidative stress predictor in a more extensive and diverse multi-ethnic population using the same ML techniques. Third, in this study, we excluded those subjects with missing observations rather than replacing missing values with imputation methods. This approach may induce the problem of reducing statistical power or yielding biased estimates if in traditional epidemiology and clinical research [52]. Therefore, we performed multiple imputation using the partially observed cases by chained equations in R package mice [53]. Although the number of selected features increased from 16 to 22 after imputation, most of them were duplicated with the model presented here. In addition, the diagnostic performance was compatible with each other (Figure S2). Last, the primary aim of creating the composite predictor was to facilitate identifying an individual's health condition and predicting longitudinal outcomes of an individual's health risks. However, by using a cross-sectional dataset, our model allowed us to develop only a diagnostic model for distinguishing oxidative stress levels of each individual. We may need a population-based longitudinal cohort dataset that follows individuals and tracks the progression of chronic diseases related to oxidative stress in order to develop a prognostic risk model and use it as a practical reference for subsequent personalized health decisions. The prognostic model is likely to be conveyable to the general public if appropriately validated.

5. Conclusions

The proposed ML algorithm is based on the cross-sectional data obtained from the Boramae medical center cohort in Korea. It allowed the development of a composite diagnostic model for oxidative stress at the individual level. The resulting predictor comprised 16 features and was proved to have excellent performance in quantifying oxidative stress risks. Considering the importance of these findings in the context of precision nutrition, the limitations of this study create opportunities for further research to enhance and expand our current understandings. The present study is the first to report a feasible approach for stratifying oxidative stress risks in a healthy population, providing appropriate strategies to engage in proactive health management to prevent diet- and lifestyle-related chronic diseases.

6. Patents

The application numbers of the patent related to this work are 10-2020-0071809 and PCT/KR2021/007312.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/antiox10071132/s1, Figure S1, Classification of oxidative stress-related disease; Figure S2, Analysis of missing data; and Table S1, Comparison of healthy controls in the training set and those in the hold-out set.

Author Contributions: Conceptualization, O.K. and J.B.; methodology and statistical analysis, Y.K. (Youjin Kim); Y.K. (Yunsoo Kim) and T.J.v.d.B.; resources, B.O.; formal analysis, Y.K. (Yunsoo Kim) and J.H.; visualization, Y.K. (Yunsoo Kim); writing—original draft preparation, Y.K. (Youjin Kim). and Y.K. (Yunsoo Kim); writing—review and editing, S.W., J.Y.K., J.B. and O.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Bio-Synergy Research Project, grant number NRF-2012M3A9C4048761 funded by the Government of South Korea (Ministry of Science, ICT, and Future Planning).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Seoul National University Boramae Medical Center (protocol code 20140929/26-2014-118/102 and date of approval 20141202) and Ewha Womans University (protocol code 86-8 and date of approval 20150102).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are included in the article.

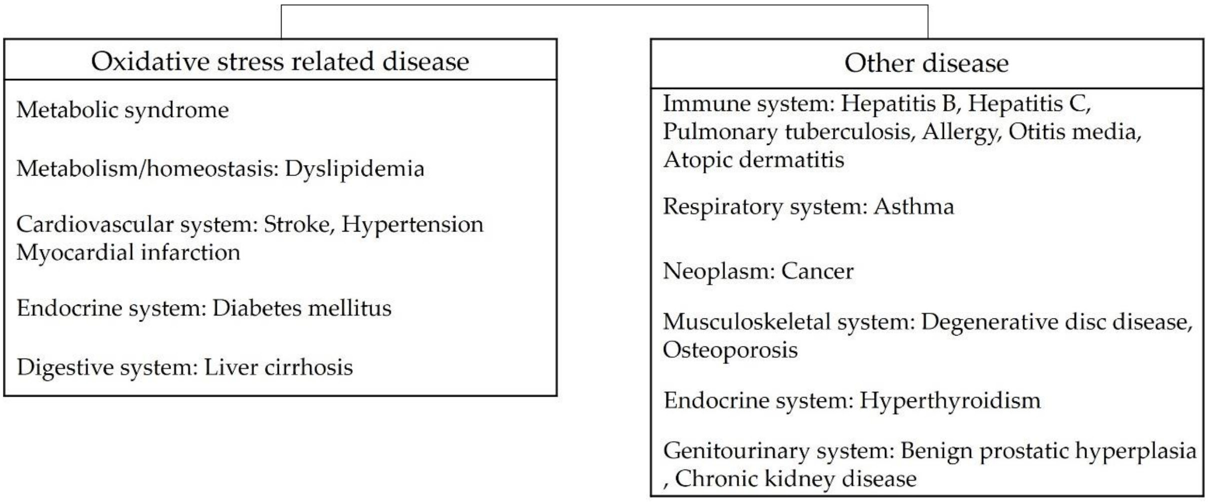
Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bonomini, F.; Rodella, L.F.; Rezzani, R. Metabolic syndrome, aging and involvement of oxidative stress. *Aging Dis.* **2015**, *6*, 109–120. [CrossRef]
2. Campbell, A.; Solaimani, P. Oxidative and inflammatory pathways in age-related chronic disease processes. In *Inflammation, Aging, and Oxidative Stress*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 95–106.
3. Ruiz-Núñez, B.; Pruimboom, L.; Dijck-Brouwer, D.A.; Muskiet, F.A. Lifestyle and nutritional imbalances associated with Western diseases: Causes and consequences of chronic systemic low-grade inflammation in an evolutionary context. *J. Nutr. Biochem.* **2013**, *24*, 1183–1201. [CrossRef]
4. Hanson, M.A.; Gluckman, P.D. Developmental origins of health and disease–global public health implications. *Best Pract. Res. Clin. Obstet. Gynaecol.* **2015**, *29*, 24–31. [CrossRef]
5. Bloomer, R.J. Decreased blood antioxidant capacity and increased lipid peroxidation in young cigarette smokers compared to nonsmokers: Impact of dietary intake. *Nutr. J.* **2007**, *6*, 39. [CrossRef]
6. Nielsen, F.; Mikkelsen, B.B.; Nielsen, J.B.; Andersen, H.R.; Grandjean, P. Plasma malondialdehyde as biomarker for oxidative stress: Reference interval and effects of life-style factors. *Clin. Chem.* **1997**, *43*, 1209–1214. [CrossRef]
7. Cherubini, A.; Ruggiero, C.; Polidori, M.C.; Mecocci, P. Potential markers of oxidative stress in stroke. *Free Radic. Biol. Med.* **2005**, *39*, 841–852. [CrossRef]
8. Furukawa, S.; Fujita, T.; Shimabukuro, M.; Iwaki, M.; Yamada, Y.; Nakajima, Y.; Nakayama, O.; Makishima, M.; Matsuda, M.; Shimomura, I. Increased oxidative stress in obesity and its impact on metabolic syndrome. *J. Clin. Investig.* **2004**, *114*, 1752–1761. [CrossRef]
9. Matsuda, M.; Shimomura, I. Increased oxidative stress in obesity: Implications for metabolic syndrome, diabetes, hypertension, dyslipidemia, atherosclerosis, and cancer. *Obes. Res. Clin. Pract.* **2013**, *7*, e330–e341. [CrossRef]
10. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Br. J. Surg.* **2015**, *102*, 148–158. [CrossRef] [PubMed]
11. Califf, R.M. Biomarker definitions and their applications. *Exp. Biol. Med.* **2018**, *243*, 213–221. [CrossRef]
12. Elmasry, G.; Kamruzzaman, M.; Sun, D.W.; Allen, P. Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: A review. *Crit. Rev. Food Sci. Nutr.* **2012**, *52*, 999–1023. [CrossRef]
13. Bouwman, J.; Vogels, J.T.; Wopereis, S.; Rubingh, C.M.; Bijlsma, S.; Ommen, B. Visualization and identification of health space, based on personalized molecular phenotype and treatment response to relevant underlying biological processes. *BMC Med. Genom.* **2012**, *5*, 1. [CrossRef]
14. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef]
15. Murdoch, T.B.; Detsky, A.S. The inevitable application of big data to health care. *JAMA* **2013**, *309*, 1351–1352. [CrossRef]
16. Dilsizian, S.E.; Siegel, E.L. Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr. Cardiol. Rep.* **2014**, *16*, 441. [CrossRef] [PubMed]
17. Kim, J.Y.; Yang, Y.J.; Yang, Y.K.; Oh, S.Y.; Hong, Y.C.; Lee, E.K.; Kwon, O. Diet quality scores and oxidative stress in Korean adults. *Eur. J. Clin. Nutr.* **2011**, *65*, 1271–1278. [CrossRef]
18. Oh, J.Y.; Yang, Y.J.; Kim, B.S.; Kang, J.H. Validity and reliability of Korean version of International Physical Activity Questionnaire (IPAQ) short form. *J. Korean Acad. Fam. Med.* **2007**, *28*, 532–541.
19. Agarwal, R.; Chase, S.D. Rapid, fluorimetric-liquid chromatographic determination of malondialdehyde in biological samples. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2002**, *775*, 121–126. [CrossRef]
20. Khoschsorur, G.; Winklhofer-Roob, B.; Rabl, H.; Auer, T.; Peng, Z.; Schaur, R. Evaluation of a sensitive HPLC method for the determination of malondialdehyde, and application of the method to different biological materials. *Chromatographia* **2000**, *52*, 181–184. [CrossRef]

21. Arsova-Sarafinovska, Z.; Aydin, A.; Sayal, A.; Eken, A.; Erdem, O.; Savaşer, A.; Erten, K.; Özgök, Y.; Dimovski, A. Rapid and simple determination of plasma and erythrocyte MDA levels in prostate cancer patients by a validated HPLC method. *J. Liq. Chromatogr. Relat. Technol.* **2007**, *30*, 2435–2444. [CrossRef]
22. Cipierre, C.; Haÿs, S.; Maucort-Boulch, D.; Steghens, J.-P.; Picaud, J.-C. Malondialdehyde adduct to hemoglobin: A new marker of oxidative stress suitable for full-term and preterm neonates. *Oxidative Med. Cell. Longev.* **2013**, *2013*, 1–6. [CrossRef]
23. Maritim, A.C.; Sanders, R.A.; Watkins, J.B., 3rd. Diabetes, oxidative stress, and antioxidants: A review. *J. Biochem. Mol. Toxicol.* **2003**, *17*, 24–38. [CrossRef]
24. Natarajan, S.K.; Thomas, S.; Ramamoorthy, P.; Basivireddy, J.; Pulimood, A.B.; Ramachandran, A.; Balasubramanian, K.A. Oxidative stress in the development of liver cirrhosis: A comparison of two different experimental models. *J. Gastroenterol. Hepatol.* **2006**, *21*, 947–957. [CrossRef]
25. Kohler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L.C.; Lewis-Smith, D.; Vasilevsky, N.A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A.M.; et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **2021**, *49*, D1207–D1217. [CrossRef]
26. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **2011**, *39*, 1–13. [CrossRef]
27. Cohen, Z.D.; Kim, T.T.; Van, H.L.; Dekker, J.J.M.; Driessen, E. A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychother. Res.* **2020**, *30*, 137–150. [CrossRef]
28. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1. [CrossRef]
29. Steyerberg, E.W.; Bleeker, S.E.; Moll, H.A.; Grobbee, D.E.; Moons, K.G. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *J. Clin. Epidemiol.* **2003**, *56*, 441–447. [CrossRef]
30. Melkumova, L.E.; Shatskikh, S.Y. Comparing Ridge and LASSO estimators for data analysis. *Procedia Eng.* **2017**, *201*, 746–755. [CrossRef]
31. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
32. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Muller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef]
33. Bolboacă, S.D. Medical diagnostic tests: A review of test anatomy, phases, and statistical treatment of data. *Comput. Math. Methods Med.* **2019**, *2019*, 1–22. [CrossRef]
34. D'Agostino, R.B.; Nam, B.-H. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures. In *Advances in Survival Analysis*; Elsevier: Amsterdam, The Netherlands, 2003; pp. 1–25.
35. Team, R.C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013; ISBN 3-900051-07-0.
36. Moons, K.G.; Altman, D.G.; Vergouwe, Y.; Royston, P. Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ* **2009**, *338*, b606. [CrossRef]
37. Block, G.; Dietrich, M.; Norkus, E.P.; Morrow, J.D.; Hudes, M.; Caan, B.; Packer, L. Factors associated with oxidative stress in human populations. *Am. J. Epidemiol.* **2002**, *156*, 274–285. [CrossRef]
38. Yamada, J.; Tomiyama, H.; Yambe, M.; Koji, Y.; Motobe, K.; Shiina, K.; Yamamoto, Y.; Yamashina, A. Elevated serum levels of alanine aminotransferase and gamma glutamyltransferase are markers of inflammation and oxidative stress independent of the metabolic syndrome. *Atherosclerosis* **2006**, *189*, 198–205. [CrossRef]
39. Waggiallah, H.; Alzohairy, M. The effect of oxidative stress on human red cells glutathione peroxidase, glutathione reductase level, and prevalence of anemia among diabetics. *N. Am. J. Med. Sci.* **2011**, *3*, 344–347. [CrossRef]
40. Widmer, C.C.; Pereira, C.P.; Gehrig, P.; Vallelian, F.; Schoedon, G.; Buehler, P.W.; Schaer, D.J. Hemoglobin can attenuate hydrogen peroxide-induced oxidative stress by acting as an antioxidative peroxidase. *Antioxid. Redox Signal.* **2010**, *12*, 185–198. [CrossRef]
41. Roehrs, M.; Conte, L.; da Silva, D.T.; Duarte, T.; Maurer, L.H.; de Carvalho, J.A.M.; Moresco, R.N.; Somacal, S.; Emanuelli, T. Annatto carotenoids attenuate oxidative stress and inflammatory response after high-calorie meal in healthy subjects. *Food Res. Int.* **2017**, *100*, 771–779. [CrossRef] [PubMed]
42. Romero-Rosales, B.L.; Tamez-Pena, J.G.; Nicolini, H.; Moreno-Treviño, M.G.; Trevino, V. Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. *PLoS ONE* **2020**, *15*, e0232103. [CrossRef] [PubMed]
43. Ford, E.; Carroll, J.A.; Smith, H.E.; Scott, D.; Cassell, J.A. Extracting information from the text of electronic medical records to improve case detection: A systematic review. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 1007–1015. [CrossRef]
44. Wu, P.Y.; Cheng, C.W.; Kaddi, C.D.; Venugopalan, J.; Hoffman, R.; Wang, M.D. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 263–273. [CrossRef]
45. Friedman, J.; Hastie, T.; Tibshirani, R.J.A.o.s. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.* **2000**, *28*, 337–407. [CrossRef]
46. Dhillon, V.; Thomas, P.; Fenech, M. Effect of common polymorphisms in folate uptake and metabolism genes on frequency of micronucleated lymphocytes in a South Australian cohort. *Mutat. Res.* **2009**, *665*, 1–6. [CrossRef]

47. Winham, S.J.; Slater, A.J.; Motsinger-Reif, A.A. A comparison of internal validation techniques for multifactor dimensionality reduction. *BMC Bioinform.* **2010**, *11*, 394. [CrossRef]
48. Steyerberg, E.W.; Harrell, F.E.; Borsboom, G.J.; Eijkemans, M.J.; Vergouwe, Y.; Habbema, J.D. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **2001**, *54*, 774–781. [CrossRef]
49. Carter, J.V.; Pan, J.; Rai, S.N.; Galandiuk, S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery* **2016**, *159*, 1638–1645. [CrossRef]
50. Park, S.; Yoo, H.J.; Jee, S.H.; Lee, J.H.; Kim, M. Weighting approaches for a genetic risk score and an oxidative stress score for predicting the incidence of obesity. *Diabetes Metab. Res. Rev.* **2019**, e3230. [CrossRef]
51. Vassalle, C.; Pratali, L.; Boni, C.; Mercuri, A.; Ndreu, R. An oxidative stress score as a combined measure of the pro-oxidant and anti-oxidant counterparts in patients with coronary artery disease. *Clin. Biochem.* **2008**, *41*, 1162–1167. [CrossRef]
52. Groenwold, R.H.; White, I.R.; Donders, A.R.; Carpenter, J.R.; Altman, D.G.; Moons, K.G. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *CMAJ* **2012**, *184*, 1265–1269. [CrossRef]
53. Zhang, Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann. Transl. Med.* **2016**, *4*, 30. [CrossRef]



Oxidative stress related disease

Metabolic syndrome

Metabolism/homeostasis: Dyslipidemia

Cardiovascular system: Stroke, Hypertension Myocardial infarction

Endocrine system: Diabetes mellitus

Digestive system: Liver cirrhosis

Other disease

Immune system: Hepatitis B, Hepatitis C, Pulmonary tuberculosis, Allergy, Otitis media, Atopic dermatitis

Respiratory system: Asthma

Neoplasm: Cancer

Musculoskeletal system: Degenerative disc disease, Osteoporosis

Endocrine system: Hyperthyroidism

Genitourinary system: Benign prostatic hyperplasia , Chronic kidney disease

Figure S1. Classification of oxidative stress-related disease

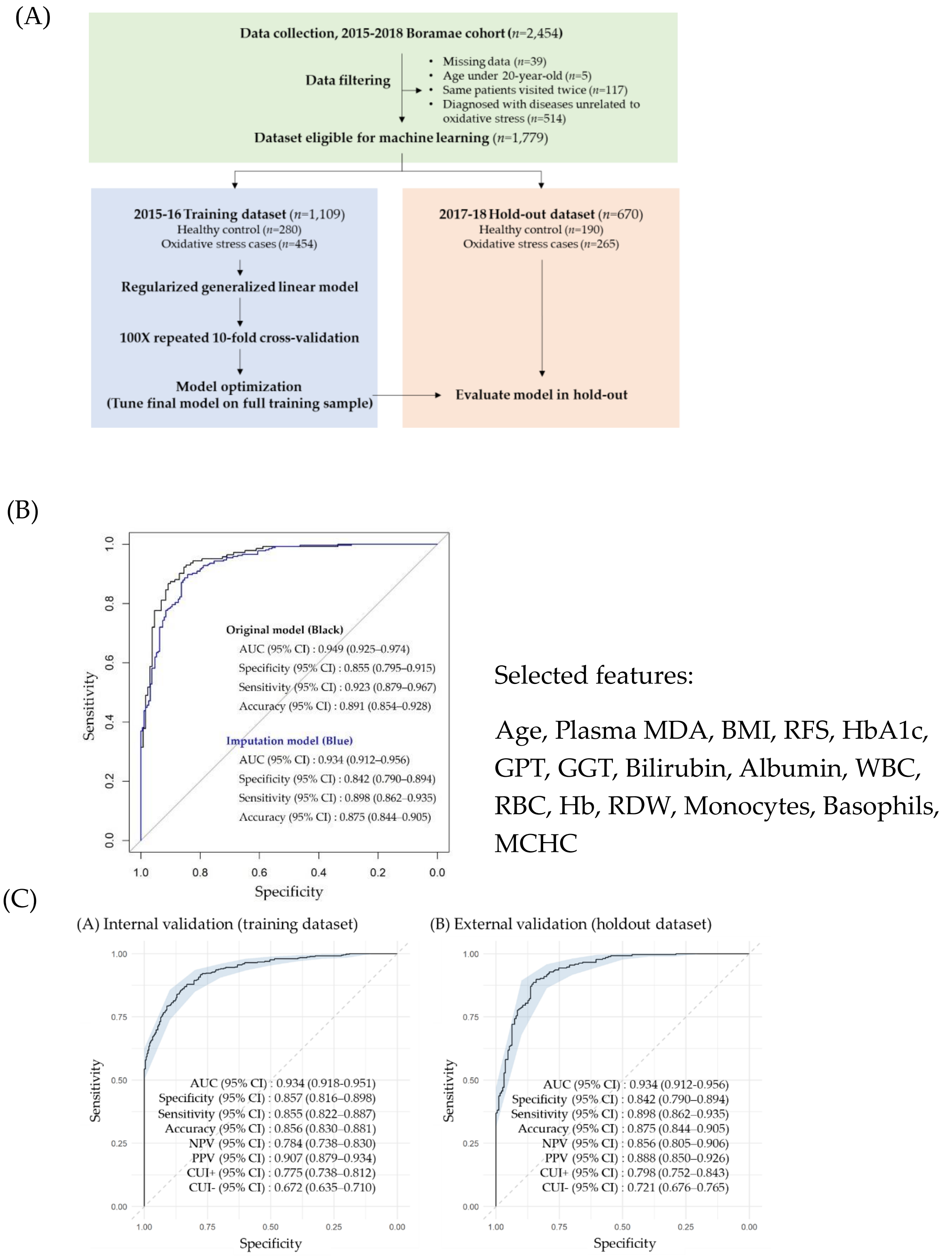


Figure S2. Analysis of missing data. (A) the schematic of the analysis, (B) selected features and performance test using receiver operating characteristic curve of best performing model, compared with the original model (*p*=0.376), and (C) internal and external validation. The diagonal line represents the reference line of 0.5. AUC, area under the curve; CI, confidence interval.

Table S1. Comparison of healthy controls in the training set and those in the hold-out set.

Features	**Normal range**	**Healthy controls**		
		2015–2016 **Training set** **(*n* = 248)**	**2017–2018** **Hold-out set** **(*n* = 131)**	***p*-value**
General characteristics (10)				
Age (years)	-	42.7 ± 10.4	38.3 ± 10.0	0.0001
Sex (males; n, %)	-	101 (40.7)	47 (35.9)	0.358
Current smoker *n* (%)	-	23 (9.3)	5 (3.8)	0.053
Smoking duration (years)	-	4.5 ± 9.0	3.1 ± 6.8	0.091
Smoking pack-year	-	3.3 ± 7.6	2.4 ± 6.4	0.215
Current drinker *n* (%)	-	145 (58.5)	78 (59.5)	0.840
Recommended food score	-	21.6 ± 8.3	20.5 ± 8.7	0.215
Physical activity (min/day)	-	128.5 ± 120.1	119.2 ± 113.8	0.465
Body mass index (kg/m^2)		21.5 ± 2.2	21.2 ± 2.2	0.261
Body fat (%)		25.5 ± 6.4	26.0 ± 6.4	0.417
Biochemical characteristics (14)				
Albumin (g/dL)	3.5-5.2	4.3 ± 0.2	4.4 ± 0.2	0.001
Alkaline phosphatase (IU/L)	44-147	62.5 ± 17.8	62.3 ± 17.2	0.908
Bilirubin (mg/dL)	0.1-1.2	1.2 ± 0.5	1.1 ± 0.4	0.101
Blood urea nitrogen (mg/dL)	8-23	12.8 ± 3.5	11.6 ± 2.8	0.0004
Creatinine (μmol/L)	Male: 61.9-114.9 Female: 53-97.2	68.2 ± 14.1	66.1 ± 12.6	0.154
C-reactive protein (mg/dL)	0-3	0.1 ± 0.2	0.1 ± 0.3	0.787
γ-Glutamyl transferase (IU/L)	0-30	18.4 ± 15.0	17.7 ± 13.7	0.662
GOT (IU/L)	Male 0-40 Female 0-32	22.6 ± 6.8	26.7 ± 56.0	0.405
GPT (IU/L)	Male 0-41 Female 0-33	18.9 ± 12.0	18.6 ± 12.3	0.810
Glycosylated hemoglobin (%)	0-5.7	5.4 ± 0.3	5.3 ± 0.3	0.008
LDL-C (mmol/L)	0-3.5	3.0 ± 0.7	3.0 ± 0.7	0.971
Total cholesterol (mmol/L)	0-5.18	5.0 ± 0.8	5.0 ± 0.8	0.457
Total protein (g/dL)	6.4-8.3	7.1 ± 0.4	7.2 ± 0.3	0.002
Uric acid (mg/dL)	Male: 2.4-6.0 Female: 3.4-7.0	4.8 ± 1.2	4.7 ± 1.1	0.691
Complete blood count data (16)				
Basophils (%)	0.0-2.0	0.4 ± 0.3	0.5 ± 0.3	0.154
Eosinophils (%)	1.0-6.0	2.6 ± 2.2	2.7 ± 2.0	0.800
ESR (mm/h)	Male: 0-22 Female: 0-29	9.0 ± 8.2	9.4 ± 11.9	0.731
Hematocrit (%)	Male: 39-52 Female: 36-48	41.4 ± 4.2	42.7 ± 3.7	0.003
Hemoglobin (g/dL)	Male: 13-17 Female: 12-16	13.9 ± 1.6	14.2 ± 1.5	0.099
Lymphocytes (%)	17.0-46.0	35.3 ± 8.8	36.4 ± 7.9	0.224
MCH (pg)	Male: 27-33 Female: 26-32	30.3 ± 1.9	30.0 ± 1.9	0.208
MCHC (g/dL)	30.0-35.0	33.5 ± 1.0	33.2 ± 0.9	0.001
MCV (fL)	Male: 81-96 Female: 79-95	90.4 ± 4.4	90.6 ± 4.7	0.647
Monocytes (%)	2.0-8.0	5.7 ± 1.5	5.8 ± 1.4	0.736
Neutrophils (%)	38.0-78.0	56.0 ± 9.7	54.7 ± 8.4	0.201

Platelets (×10³/μL)	130-400	249.1	±	51.9	258.5	±	54.0	0.101
Platelet distribution width (%)	8.3-56.6	11.8	±	1.6	11.8	±	1.5	0.734
Red blood cell count (×10⁶/μL)	Male: 4.2-6.3 Female: 4.0-5.4	4.6	±	0.4	4.7	±	0.4	0.004
RDW (%)	Male: 12.2-16.1 Female: 11.8-14.5	13.1	±	1.0	13.1	±	1.1	0.998
White blood cell count (×103/μL)	4.0-10.0	5.2	±	1.3	5.2	±	1.3	0.959
High-performance liquid chromatography analysis data (3)								
Plasma MDA (mM)	-	4.4	±	2.4	4.3	±	2.2	0.622
Erythrocyte MDA (mM)	-	6.1	±	6.2	8.8	±	8.6	0.002
Urine MDA (mM)	-	2.9	±	1.8	2.7	±	1.7	0.369

Values are mean ± SD or number (percentage). Differences between the groups were compared using Student's *t*-test for continuous variables and Chi-square tests for categorical variables. AP, alkaline phosphatase; BMI, body mass index; BUN, blood urea nitrogen; CRP, C-reactive protein; GOT, glutamate oxaloacetate transaminase; GPT, glutamate pyruvate transaminase; GT, γ-Glutamyl transferase; LDL-C, low-density lipoprotein cholesterol; ESR, erythrocyte sedimentation rate; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; MDA, malondialdehyde; PDW, platelet distribution width; RDW, red blood cell distribution width; RFS, recommended food score; TC, total cholesterol.