Establishment of Metabolic Syndrome Prediction Model for Occupational Population based on the Lasso Regression Algorithm

Qiao-Ying Xie
Hangzhou Occupational Disease Prevention and Treatment Hospital: Hangzhou Hospital for the Prevention and Treatment of Occupational Disease

Ming-Wei Wang
Hangzhou Normal University Affiliated Hospital

Zu-Ying Hu
Hangzhou Occupational Disease Prevention and Treatment Hospital: Hangzhou Hospital for the Prevention and Treatment of Occupational Disease

Yan-Ming Chu
Zhejiang Geriatric Care Hospital

Cheng-Jian Cao
Hangzhou Occupational Disease Prevention and Treatment Hospital: Hangzhou Hospital for the Prevention and Treatment of Occupational Disease

Min Wang
Jiangsu Normal University

Cong Wang
Jiangsu Normal University

Jing-Yu Kang
Jiangsu Normal University

Xin-Yan Fu
Hangzhou Normal University Affiliated Hospital

Xing-Wei Zhang
Hangzhou Normal University Affiliated Hospital

Zhan-Hui Feng
Guizhou Medical University

Jian-Bo Li
Jiangsu Normal University

Yong-Ran Cheng (chengyr@zjams.com.cn)
Hangzhou Medical College https://orcid.org/0000-0003-2646-560X

Research

Keywords: Lasso regression algorithm, metabolic syndrome, occupational population

Posted Date: December 31st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-136449/v1

Establishment of metabolic syndrome prediction model for occupational population based on the Lasso regression algorithm

Qiao-Ying Xie[1#], Ming-Wei Wang[2#], Zu-Ying Hu[1], Yan-Ming Chu[3], Cheng-Jian Cao[1], Min-Wang[4], Cong Wang[5], Jing-Yu Kang[5], Xin-Yan Fu[2], Xing-Wei Zhang[2], Zhan-Hui Feng[6*], Jian-Bo Li[5*], Yong-Ran Cheng[7*]

Affiliations:

1 Occupational disease department, Hangzhou occupational disease prevention and control hospital, Hangzhou, 310014, China;

2 Metabolic Disease Center, Affiliated Hospital of Hangzhou Normal University, Hangzhou,310015, China;

3 Zhejiang Geriatric Care Hospital ,Hangzhou, 311300, China;

4 Kewen college, Jiangsu Normal University , Xuzhou,221100, China;

5 School of mathematics and statistics, Jiangsu Normal University , Xuzhou,221100, China;

6 Neurological Department, Affiliated Hospital of Guizhou Medical University, Guiyang, China;

7 School of public health, Hangzhou Medical college, Hangzhou, 311300, China

Qiao-Ying Xie and Ming-Wei Wang contributed equally to this work

*Corresponding Author: Zhan-Hui Feng ,Jian-Bo Li and Yong-Ran Cheng

E-mail:

Qiao-Ying Xie :xieqiaoying1978@163.com;

Ming-Wei Wang: wmw990556@163.com;

Zu-Ying Hu: yinghz@163.com;

Yan-Ming Chu: chuyanming1965@163.com;

Cheng-Jian Cao: caocj2005@163.com;

Min-Wang: wangmin920118@163.com;

Cong Wang: wangcong2019@jsnu.edu.cn;

Jing-Yu Kang: 15957792738@139.com;

Xin-Yan Fu: 13735527510@163.com;

Xing-Wei Zhang: hsdzxw@126.com;

Jian-Bo Li: lijianbo@jsnu.edu.cn;

Zhan-Hui Feng: h9450203@126.com;

Yong-Ran Cheng: chengyr@zjams.com.cn

Abstract

Background: Metabolic syndrome (MS) screening is important for the early detection of occupational population. This study aimed to screen out biomarkers related to MS and establish a risk assessment and prediction model for the routine physical examination of an occupational population.

Methods: The least absolute shrinkage and selection operator (Lasso) regression algorithm of machine learning was used to screen biomarkers related to MS. Then, the accuracy of the logistic regression model was further verified based on the Lasso regression algorithm. Finally, the screened biomarkers were used to establish a logistic regression model and calculate the odds ratio (OR) of the corresponding biomarkers.

Results: A total of 2844 occupational workers were included, and 10 biomarkers related to MS were screened. The area under the curve (AUC) value for non-Lasso and Lasso regression was 0.652 and 0.907, respectively. The established risk assessment model revealed that the main risk factors were basophil absolute count (OR: 3.38), platelet packed volume (OR: 2.63), leukocyte count (OR: 2.01), red blood cell count (OR: 1.99), and alanine aminotransferase level (OR: 1.53).

Conclusion: The risk assessment model based on the Lasso regression algorithm helped identify Metabolic syndrome with high accuracy in physically examining an occupational population.

Keywords: Lasso regression algorithm, metabolic syndrome, occupational population

Introduction

Metabolic syndrome (MS) refers to a group of metabolism-related diseases, including obesity, dyslipidemia, diabetes/impaired glucose tolerance, hypertension, and other diseases [1]. The number of patients with MS has increased with the increasing number of obese patients worldwide [2]. At present, the global prevalence of MS is about 25%, indicating that nearly one billion people are affected. Among these, the occupational population occupies a significant part, and still continues to increase [3]. It has posed a huge economic burden, and has become a serious public health problem.

China ranks first in the world, with nearly 900 million occupational people. Every year, nearly 25 million workers suffer from occupational work hazards, among which MS is already an important risk factor seriously affecting the health of the occupational population [4]. Many studies were conducted on the relationship between the working environment of the occupational population and MS. Ma et al. confirmed that exposure to heavy metal elements in the work environment affected the body's metabolic function and increased the risk of MS in the Chinese population [5]. Huang et al. confirmed that the long-term exposure to noise in the work environment increased the chance of suffering from MS in the Chinese professional population [6]. At the same time, some related studies confirmed the relationship of MS with the type of work in different occupational groups [7-9].

Therefore, performing early MS screening for the occupational population is of great significance. Machine learning, whereby a computer algorithm learns from prior experience, was recently shown to have better performance over traditional statistical modeling approaches [10-11]. The machine learning algorithms have been widely used to screen biomarkers for related diseases with the rapid development of artificial intelligence [12-14]. Various supervised machine learning models based on the least absolute shrinkage and selection operator (Lasso) regression algorithm have been successfully applied to medical data [15]. However, no relevant studies used the Lasso regression algorithm to screen relevant biomarkers for MS.

Therefore, the risk of MS can be better predicted if the biomarkers related to MS

are screened, and a risk prediction model is established for routine physical examination markers. In this study, the Lasso regression feature selection algorithm of machine learning was used to screen the biomarkers related to MS, and a risk prediction model was established. The objective of the study was to provide early warning and preventive measures for MS in an occupational population.

Materials and Methods

Population and data collection

This study included occupational workers with high-temperature operations in Zhejiang Province, China (referring to operations with an average wet bulb globe temperature(WBGT) index of ⩾25°C at the workplace during the production process) between September 2010 and September 2020. The working environment included the metallurgical industry, including steelmaking, ironmaking, steel rolling, coking, and so forth; casting, forging, heat treatment, and so forth in the machinery manufacturing industry; and kiln workers and furnace workers in the glass and refractory industries. A total of 3577 workers were examined, of which 733 workers were excluded due to incomplete records and errors. Finally, 2844 workers were selected for the study. This study included 32 basic biomarkers for routine physical examination in the population (Table 1).

Identification of MS

The diagnostic criteria referred to the diagnostic criteria set by the Diabetes Branch of the Chinese Medical Association for the Chinese population, and three or more of the following four components indicated MS [16]: (1) overweight or obesity: body mass index (BMI) = weight (kg)/height (m^2), BMI ⩾25; (2) hyperglycemia: fasting plasma glucose(FPG) ⩾6.1 mmol/L; (3) hypertension: systolic blood pressure(SBP)/diastolic blood pressure (DBP) ⩾140/90 mm Hg; and (4) dyslipidemia: TG ⩾1.7 mmol/L or low density lipoprotein cholesterol (HDL-C) <0.9 mmol/L (male) or <1.0mmol/L (female).

Lasso regression

Lasso regression feature selection is an unbiased estimation used to process high-dimensional complex collinearity data. The basic idea is to construct a penalty function to select the main variables with a strong correlation with the output parameters from the input variables and build a refined regression model [17]. The penalty function constructed is as follows:

$$\widehat{\beta_0}, \hat{\beta} = \arg\max\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2\}$$

$$\text{Subject to } \sum_{j=1}^{p}|\beta_j| \leq \lambda$$

where y_i is the dependent variable, $X_{ij} = (X_{i1}, X_{i1}, \ldots, X_{in})$ is an independent variable, β_j is the regression coefficient of the jth variable, and the value of λ can be [0, + ∞). Lasso feature selection compresses the model coefficients by increasing the penalty coefficient λ. When the absolute value of the regression coefficient Lasso estimate in the model is less than the absolute value of the minimum regression coefficient, some of the coefficients of the variables not strongly correlated are compressed to 0, and the variables corresponding to the coefficients with the estimated value of 0 are eliminated. In this way, the independent variables strongly related to the dependent variable are screened to achieve the purpose of feature selection.

Statistical analysis

A one-way analysis of variance was used to compare the differences between the metabolome and non-metabolome biomarkers in routine physical examination. The random sampling method was used to deal with the sample imbalance between workers with and without MS [18]. The tenfold cross-validation method was used to determine the best Lasso penalty coefficient. Based on the selected biomarkers, the logistic regression risk prediction model was established, and the value of each biomarker was given. A test *P* value less than 0.05 indicated a statistically significant difference. The Lasso algorithm used the "glmnet" package for calculation. The receiver operating characteristic (ROC) curve was used to evaluate the accuracy of the predictive risk

model. All analyses were performed using the statistical programming environment R (version 3.6.0).

Results

Basic characteristics of the population and biomarkers in physical examination

A total of 2844 occupational workers were involved (Table 2), including 655 with MS (638 men and 17 women) and 2189 without MS (1936 men and 253 women). The body weight was greater in the MS group (78.4 kg) than in the non-MS group (64.9 kg). The average systolic blood pressure was higher in the MS group (86.5/154.1 mm Hg) than in the non-MS group (72.5/118.5 mm Hg). The one-way analysis of variance revealed differences in the expression of 13 physical examination markers ($P < 0.05$) (Table 3).

Selection of physical examination biomarkers

The biomarkers were selected using the Lasso binary logistic regression model. (Figure. 1A). The tuning parameter (λ) selection in the Lasso model used tenfold cross-validation based on the minimum criteria. The area under the binomial deviance curve was plotted versus log (λ). Dotted vertical lines were drawn at the optimal values using the minimum criteria and the 1 standard error of the minimum criteria (the 1-SE criteria). Further, log (λ) = –4.331 was chosen (1-SE criteria) according to tenfold cross-validation of Lasso coefficient profiles of the 32 features. A coefficient profile plot was produced against the log (λ) sequence (Figure. 1B). A vertical line was drawn at the value selected using tenfold cross-validation, where optimal λ resulted in 10 nonzero coefficients. Finally, the physical examination markers related to MS were selected (Figure. 1C).

Establishment of a risk prediction model

A multiple logistic regression model was established. The accuracy of the model was compared. All 32 physical examination markers were incorporated into the model. The predicted results of the model are shown in Figure 2A, indicating that the AUC of

the model was 0.652. The prediction result of the model after incorporating the final 10 markers into the model is shown in Figure 2B. The AUC of the model was 0.907. A multiple logistic regression model was established using the 10 physical examination markers selected; the analysis results are shown in Figure 3. Only two physical examination marker showed no statistical significance in the prediction model ($P > 0.05$). The first five risk factors were the basophils absolute value (OR: 3.38), platelet packed count (OR: 2.63), and leukocyte count (OR: 2.01), red blood cell count (OR: 1.99), and alanine aminotransferase level (OR: 1.53).

Discussion

The health of the occupational population has a strong relationship with the working environment. This population has high work pressure, disordered work and rest, irregular diet, and lack of exercise. These inevitable adverse factors increase the risk of MS. This study proposed the potential predictors based on the screening of routine physical examination biomarkers; the established MS prediction model could be extended to clinical and physical examination centers. The routine physical examination data were used to conduct an early risk assessment of MS in the occupational population.

Hsiao and Yang conducted a 2-year (2003–2005) and 5-year (1997–2006) follow-up on a Chinese population [19-20]. They both confirmed the routine examination of biomarkers such as serum cholesterol, triglyceride, and blood glucose levels, height, weight, blood pressure, and so forth. The multivariate logistic regression analysis could be used as an effective predictor of MS. In this study, 10 biomarkers related to MS were further screened, including red blood cell count, total protein level, percentage of neutrophils, red blood cell distribution width CV, absolute number of neutrophils, leukocyte count, absolute value of basophils, alanine aminotransferase level, monocyte count, and platelet count. These potential biomarkers could be used to assess the risk of MS.

A low-level inflammatory state is considered to be a major potential mechanism of MS. Recent studies have found that the leukocyte count is associated with MS and

cardiovascular disease. A longitudinal cohort study of a healthy population in China showed a significant correlation between white blood cell count and MS (relative risk = 2.66). At the same time, the total numbers of white blood cell, neutrophils, monocytes, and basophils were the risk factors for obesity [21]. Liu et al. found a significant positive correlation between alanine aminotransferase level and risk of MS through quantitative and qualitative analyses, which had a predictive value for the incidence of MS [22]. Further, a positive correlation was reported between red blood cell parameters, hematocrit, and MS for a large longitudinal cohort in China [23]. Laufer et al. found that the prevalence of MS was 29% when the red blood cell distribution width was less than 14%, and the prevalence of MS was 34% when the red blood cell distribution width was more than 14% [24]. The findings on the biomarkers screened in the aforementioned studies were the same as those in the present study.

In this study, the Lasso feature selection algorithm was used to accurately screen the physical examination markers related to MS for an occupational population. The study showed that Lasso feature selection made the screened biomarkers more explanatory and reduced the complexity of the subsequent risk model. Other machine learning algorithms, such as decision trees [25], random forests [26], neural networks [27], and so forth, can be used to compare the accuracy of each method in future studies.

Conclusions

This study selected 10 physical examination indicators related to MS based on the Lasso algorithm in machine learning. An accurate risk prediction model for MS was established. The use of common indicators and examination items in the health examination of ordinary occupational populations provides a basis for the use of cheap and portable methods to realize the risk prediction of MS.

References

1. Samson SL, Garber AJ. Metabolic syndrome. Endocrinol Metab Clin North Am. 2014 Mar;43(1):1-23.
2. Saklayen MG. The Global Epidemic of the Metabolic Syndrome. Curr Hypertens Rep. 2018 Feb 26;20(2):12.
3. van Zon SKR, Amick Iii BC, de Jong T, Brouwer S, Bültmann U. Occupational distribution of metabolic syndrome prevalence and incidence differs by sex and is not explained by age and health behavior: results from 75 000 Dutch workers from 40 occupational groups. BMJ Open Diabetes Res Care. 2020 Jul;8(1):e001436.
4. Zhang, Zhansai et al. "China's occupational health challenges." *Occupational medicine (Oxford, England)* vol. 67,2 (2017): 87-90.
5. Ma J, Zhou Y, Wang D, Guo Y, Wang B, Xu Y, Chen W. Associations between essential metals exposure and metabolic syndrome (MetS): Exploring the mediating role of systemic inflammation in a general Chinese population. Environ Int. 2020 Jul;140:105802.
6. Huang T, Chan TC, Huang YJ, Pan WC. The Association between Noise Exposure and Metabolic Syndrome: A Longitudinal Cohort Study in Taiwan. Int J Environ Res Public Health. 2020 Jun 14;17(12):4236.
7. Zhang B, Pan B, Zhao X, Fu Y, Li X, Yang A, Li Q, Dong J, Nie J, Yang J. The interaction effects of smoking and polycyclic aromatic hydrocarbons exposure on the prevalence of metabolic syndrome in coke oven workers. Chemosphere. 2020 May;247:125880.
8. Tsai TY, Cheng JF, Lai YM. Prevalence of metabolic syndrome and related factors in Taiwanese high-tech industry workers. Clinics (Sao Paulo). 2011;66(9):1531-5.
9. Lin CY, Lin CM. Occupational Assessments of Risk Factors for Cardiovascular Diseases in Labors: An Application of Metabolic Syndrome Scoring Index. Int J Environ Res Public Health. 2020 Oct 16;17(20):7539.
10. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018 Apr 3;319(13):1317-1318.
11. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med. 2017 Jun 29;376(26):2507-2509.
12. Booth TC, Williams M, Luis A, Cardoso J, Ashkan K, Shuaib H. Machine learning and glioma imaging biomarkers. Clin Radiol. 2020 Jan;75(1):20-32..
13. Boissoneault J, Sevel L, Letzen J, Robinson M, Staud R. Biomarkers for Musculoskeletal Pain Conditions: Use of Brain Imaging and Machine Learning. Curr Rheumatol Rep. 2017 Jan;19(1):5.
14. Radhakrishnan A, Damodaran K, Soylemezoglu AC, Uhler C, Shivashankar GV. Machine Learning for Nuclear Mechano-Morphometric Biomarkers in Cancer Diagnosis. Sci Rep. 2017 Dec 20;7(1):17946.
15. Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, Ma ZL, Liu ZY. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. J Clin Oncol. 2016 Jun 20;34(18):2157-64.
16. Alberti KG, Zimmet P, Shaw J. Metabolic syndrome--a new world-wide definition. A Consensus Statement from the International Diabetes Federation. Diabet Med. 2006

May;23(5):469-80.
17. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007 Dec 30;26(30):5512-28.
18. Chetchotsak D, Pattanapairoj S, Arnonkijpanich B. Integrating new data balancing technique with committee networks for imbalanced data: GRSOM approach. Cogn Neurodyn. 2015 Dec;9(6):627-38.
19. Hsiao FC, Wu CZ, Hsieh CH, He CT, Hung YJ, Pei D. Chinese metabolic syndrome risk score. South Med J. 2009 Feb;102(2):159-64.
20. Yang XH, Tao QS, Sun F, Cao CK, Zhan SY. [Setting up a risk prediction model on metabolic syndrome among 35-74 year-olds based on the Taiwan MJ Health-checkup Database]. Zhonghua Liu Xing Bing Xue Za Zhi. 2013 Sep;34(9):874-8.
21. Meng W, Zhang C, Zhang Q, Song X, Lin H, Zhang D, Zhang Y, Zhu Z, Wu S, Liu Y, Tang F, Yang X, Xue F. Association between leukocyte and metabolic syndrome in urban Han Chinese: a longitudinal cohort study. PLoS One. 2012;7(11):e49875.
22. Liu CF, Zhou WN, Fang NY. Gamma-glutamyltransferase levels and risk of metabolic syndrome: a meta-analysis of prospective cohort studies. Int J Clin Pract. 2012 Jul;66(7):692-8.
23. Hwang HJ, Kim SH. Inverse relationship between fasting direct bilirubin and metabolic syndrome in Korean adults. Clin Chim Acta. 2010 Oct 9;411(19-20):1496-501.
24. Laufer Perl M, Havakuk O, Finkelstein A, Halkin A, Revivo M, Elbaz M, Herz I, Keren G, Banai S, Arbel Y. High red blood cell distribution width is associated with the metabolic syndrome. Clin Hemorheol Microcirc. 2015 Sep 25;63(1):35-43.
25. Prosperi MC, Belgrave D, Buchan I, Simpson A, Custovic A. Challenges in interpreting allergen microarrays in relation to clinical symptoms: a machine learning approach. Pediatr Allergy Immunol. 2014 Feb;25(1):71-9.
26. Zhang L, Huettmann F, Zhang X, Liu S, Sun P, Yu Z, Mi C. The use of classification and regression algorithms using the random forests method with presence-only data to model species' distribution. MethodsX. 2019 Sep 28;6:2281-2292.
27. Kriegeskorte N, Golan T. Neural network models and deep learning. Curr Biol. 2019 Apr 1;29(7):R231-R236.

Figure 1. (A) Tuning parameter (λ) selection in the Lasso model used tenfold cross-validation based on the minimum criteria. (B) Changes in 32 marker coefficients with the penalty parameter (λ). (C) 32 marker coefficients obtained according to the selected best penalty parameter (λ).
Figure 2. Receiver operating characteristic (ROC) curve with area under the curve values for (A) non-Lasso regression and (B) Lasso regression.
Figure 3. Study population for multivariate logistic regression analyses.

Table 1. Types of medical markers included in the study

Table 2. Basic characteristics of the population

Table 3. Basic characteristics of routine physical examination markers

Abbreviations

MS: Metabolic syndrome; OR: odds ratio; AUC :area under the curve; ROC: receiver operating characteristic;

Acknowledgment

We thank the physical examination center of Hangzhou occupational disease prevention and control hospital for free regular physical examination for occupational workers.

Funding

The presented study was supported by the Hangzhou Science and technology development plan projects (20140633B32;20200834M29); Youth fund of Zhejiang Academy of Medical Sciences (No.2019Y009); Medical and Technology Project of Zhejiang Province (No.2021HY127,No.2020362651,No.2021KY890); Hangzhou science and Technology Bureau fund (No.20191203B96;No.20191203B105); Clinical Research Fund of Zhejiang Medical Association(No.2020ZYC-A13); Hangzhou Health and Family Planning Technology Plan key projects（2017ZD02）

Availability of data and materials

The original data are available on request to Mrs.Qiao-Ying Xie, or at the Hangzhou occupational disease prevention and control hospital, Hangzhou, 310014, China

Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Ethics approval and consent to participate

Not applicable. My manuscript does not report on or involve the use of any animal or human data or tissue.

Consent for publication

Not applicable. All data were supplied and analysed in an anonymous format, without access to personal identifying information

Author contributions

YRC, JBL, ZHF conceived the study and designed the analysis, ZYH ,YMC and CJC curated the clinical data, MW, CW and JYK performed statistical analysis, QYX and MWW wrote the first draft of the manuscript, XYF and XWZ Participate in revision

the manuscript, and all other authors contributed to revision of the manuscript.

Figures

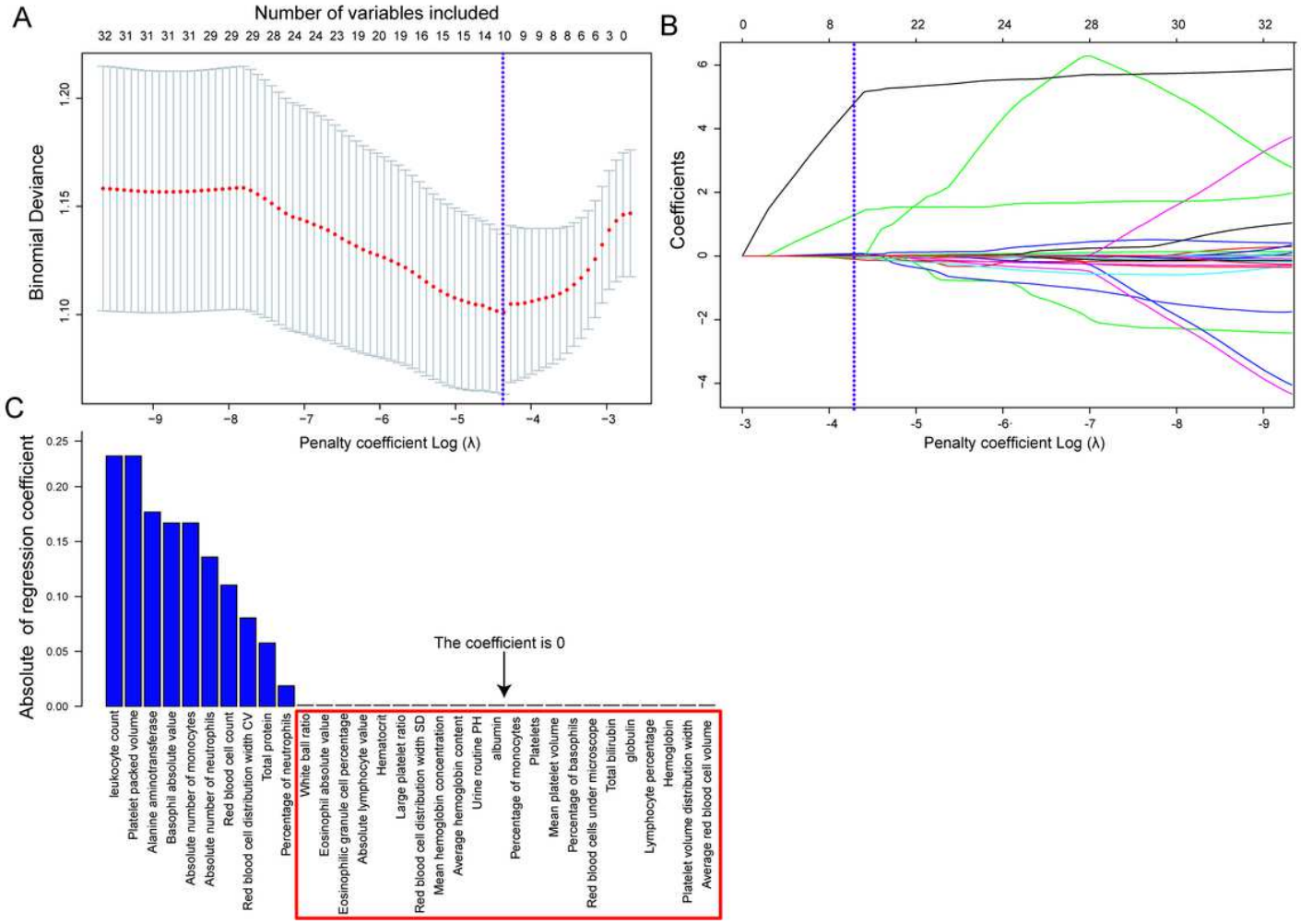

Figure 1

(A) Tuning parameter (λ) selection in the Lasso model used tenfold cross-validation based on the minimum criteria. (B) Changes in 32 marker coefficients with the penalty parameter (λ). (C) 32 marker coefficients obtained according to the selected best penalty parameter (λ).

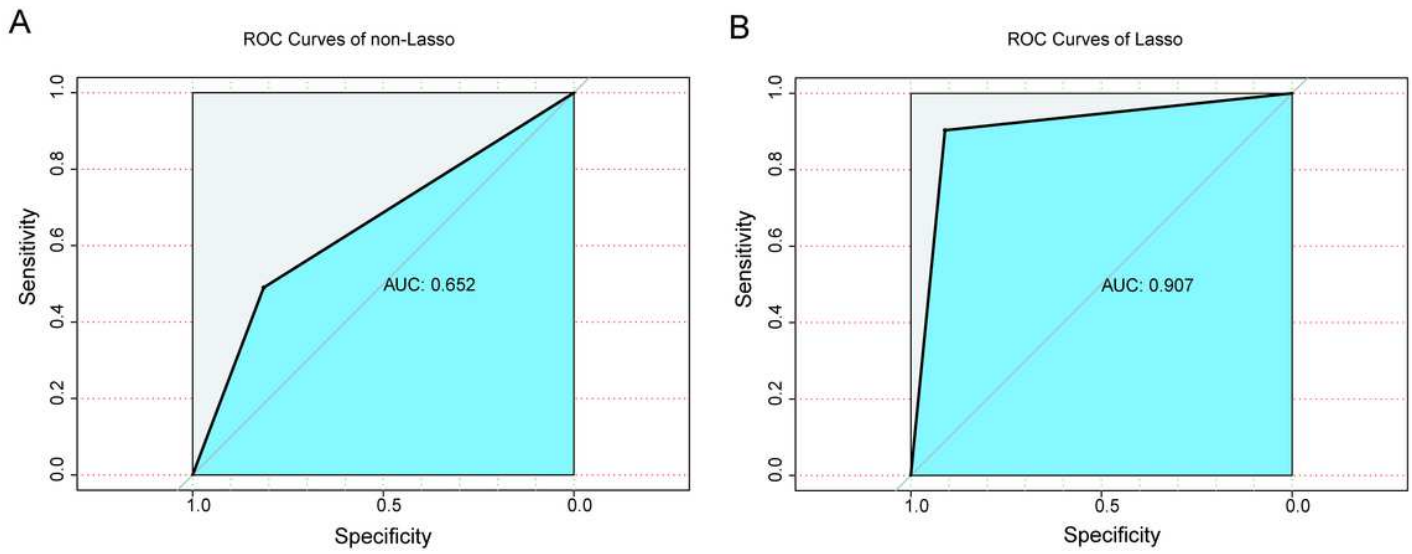

Figure 2

Receiver operating characteristic (ROC) curve with area under the curve values for (A) non-Lasso regression and (B) Lasso regression.

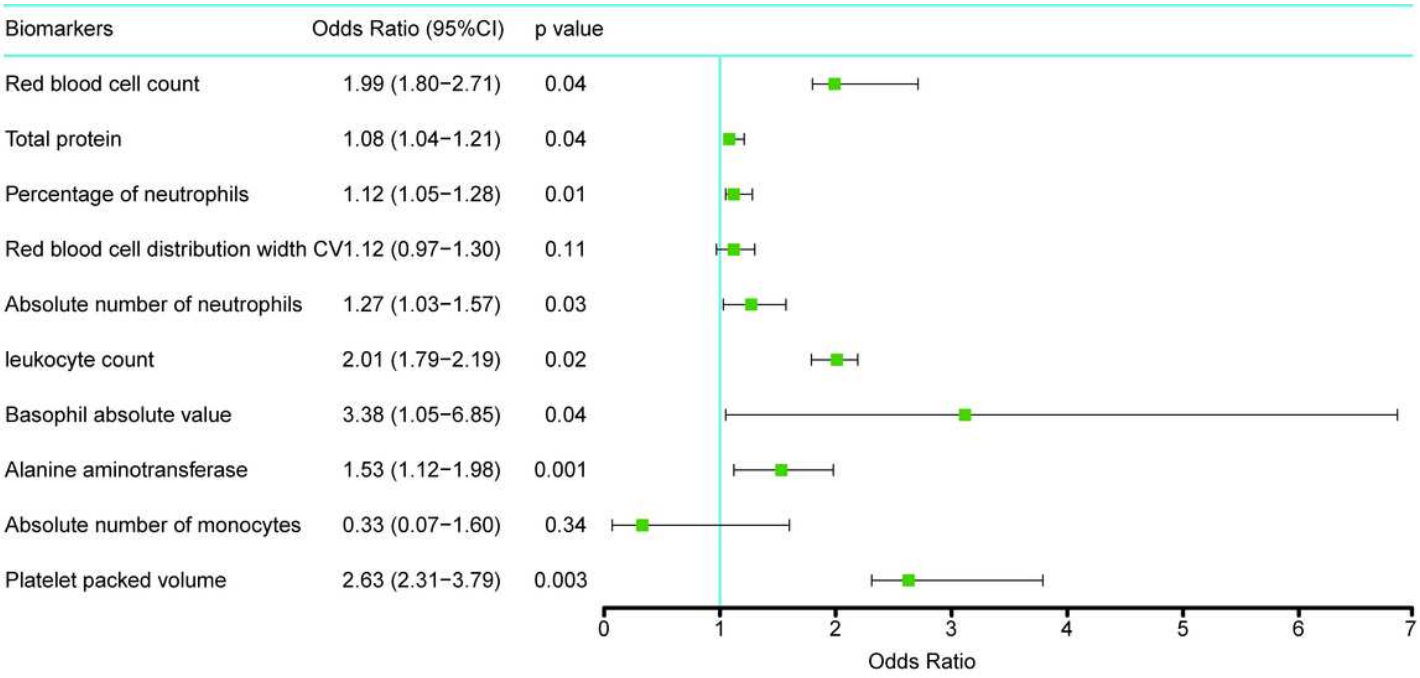

Figure 3

Study population for multivariate logistic regression analyses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx

- Table2.docx
- Table3.docx