Employing a random forest model to forecast the likelihood of coronary artery lesions in Kawasaki disease: a study centered on four biomarkers

Chu-Xiong Gong¹, Yue-Wen Li², Ya-Min Li³, Zi-Yu Wang¹, Hui-Qing Gao¹, Xiao-Mei Liu¹*concerning

¹Department of Infectious Diseases, Children's Hospital Affiliated of Kunming Medical University, Kunming Children's Hospital, Kunming 650228, China.
²Department of Rehabilitation and Palliative Medicine, The Third Affiliated Hospital of Kunming Medical University, Kunming 650000, China.
³Corresponding to: Xiao-Mei Liu, Department of Infectious Diseases, Children's Hospital Affiliated of Kunming Medical University, Kunming Children's Hospital, No. 28, Shulin Street, Xishan District, Kunming 650228, China. E-mail: 1043443230@qq.com.

Author contributions
Chu-xiong Gong and Xiao-mei Liu were the main contributors to this study and were responsible for data organization, analysis, and manuscript writing. Yuewen Li, Yamin Li, Ziyu Wang and Huiqing Gao contributed to the collection and organization of clinical data.

Competing interests
The authors declare no conflicts of interest.

Acknowledgments
This work was supported by the Joint Special Fund for Applied Basic Research of Yunnan Provincial Science and Technology Department and Kunming Medical University (20210140700001-217).

Peer review information
Medical Data Mining thanks Han-qing Zhao and anonymous reviewers for their contribution to the peer review of this paper.

Abbreviations
KD, Kawasaki disease; CAL, coronary artery lesions; WBC, white blood cell count; PLT, platelet; CRP, C-reactive protein; ALB, albumin; CKMB, creatine kinase isoenzyme MB; LASSO, Least Absolute Shrinkage and Selection Operator; AUC, area under the curve; DCA, Decision Curve Analysis; RMSE, root mean square error; NLR, neutrophil-to-lymphocyte ratio.

Citation

Executive editor: Xin-Yun Zhang.
Received: 14 January 2024; Accepted: 04 March 2024; Available online: 11 March 2024. © 2024 By Author(s). Published by TMR Publishing Group Limited. This is an open access article under the CC-BY license. (https://creativecommons.org/licenses/by/4.0/)
Introduction

Kawasaki disease (KD) is an acute immune-mediated vasculitis that primarily affects children under the age of 5 [1, 2]. KD is most commonly observed in East Asia, particularly in Japan, South Korea, and China, although its distribution is relatively equal in other countries and regions. Roughly 25% of KD patients experience coronary artery lesions (CAL). There has been a notable decrease in CAL incidence with the use of intravenous immunoglobulin. However, CAL still occurs in 3%–5% of affected children, and it is responsible for nearly all fatalities in KD cases [3–5]. Currently, the primary reliance for diagnosing KD lies in clinical symptoms such as the duration of fever, mucosal changes, and limb edema. Delaying the diagnosis and misdiagnosing, particularly in cases of atypical KD, can lead to delayed treatment during the acute phase, thereby heightening the risk of CAL. In recent years, there has been a steady rise in the incidence of CAL [2]. Swift and accurate evaluation of the risk of coronary artery damage, coupled with timely intervention, can enhance patient prognosis. As a result, it is paramount to develop a dependable predictive model to identify KD patients at high risk for CAL and to guide clinical practice effectively.

Materials and methods

Material

This study collected clinical and laboratory data from patients diagnosed with Kawasaki disease at the Children’s Hospital of Kunming Medical University between January 2016 and December 2022. This study was approved by the Ethics Committee of Kunming Children’s Hospital (2023-05-016-K01).

Inclusion criteria

In the study, the American Heart Association criteria were used to diagnose Kawasaki disease. Patients were required to have a prolonged fever lasting more than 5 days and exhibit at least four out of the following five symptoms: conjunctival injection, changes in the lips or oral cavity, extremity changes (such as peripheral edema, peripheral erythema, or periungual desquamation), rash, and cervical lymphadenopathy in order to be diagnosed with Kawasaki disease.

Echocardiography was utilized to assess CAL or aneurysm formation. A coronary artery diameter equal to or larger than 4 mm was considered indicative of CAL or aneurysm formation for patients aged 5 years and older, while a coronary artery diameter equal to or larger than 3 mm was used as the threshold for CAL or aneurysm detection in patients younger than 5 years.

Evaluated variables

When constructing the model, we actively considered several variables. These variables included demographic characteristics such as sex, ethnicity, and age in months. Additionally, we took into account clinical features such as the presence of fever, as well as laboratory test results including white blood cell count (WBC), platelet (PLT) count, red blood cell count, hemoglobin concentration, lymphocyte, neutrophil, monocyte, C-reactive protein (CRP) level, direct bilirubin level, albumin (ALB) level, alanine transaminase level, aspartate transaminase level, creatine kinase isoenzyme MB (CKMB) level, prothrombin time, and activated partial thromboplastin time. It should be emphasized that all the collected data included in this study were complete, true, and valid.

Statistical analysis

This study utilized R version 4.2.3 for all statistical analyses. Continuous variables were presented as mean ± standard deviation or median with interquartile range, while categorical variables were expressed as rates. Differences between the two groups were evaluated using independent t-tests or Mann-Whitney U tests for continuous variables, and chi-square tests or Fisher’s exact tests for categorical variables. A P < 0.05 is widely accepted as indicating statistical significance.

To identify the predictive factors associated with the occurrence of interest, the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm and Random Forest algorithm were employed [6, 7]. LASSO regularization method shrinks the coefficients to reduce model complexity. The well-known machine learning algorithm, Random Forest, has demonstrated high accuracy in disease risk prediction and diagnosis. To prevent overfitting of the model, we conduct 5-fold cross-validation on the model to determine the optimal K value, then proceed with random forest computation, and evaluate the degree of fit using R². We compared the results obtained from these machine learning methods and considered the intersection of the findings as the final set of predictive factors. Subsequently, we included the identified factors in a binary logistic regression analysis to evaluate their predictive significance and generate a regression model. We employed the ‘lrm’ package for nomogram visualization and plotted the receiver operating characteristic curve to determine the optimal cutoff value for the area under the curve (AUC). This analysis helped us assess the predictive accuracy and discriminatory ability of both individual predictive factors and the overall model. We also plotted calibration curves to evaluate the calibration of the nomograms. Furthermore, we utilized Decision Curve Analysis (DCA) to reassess the model’s accuracy [6, 7].

Results

Baseline characteristics

We included a total of 228 patients diagnosed with KD in this study based on the defined inclusion criteria and availability of complete laboratory results. Among the included patients, 141 (61.84%) were male and 87 (38.16%) were female. The majority of patients belonged to the Han ethnicity (88.16%), while 27 (11.84%) came from minority ethnic groups. The mean age of the patients was 42.28 ± 28.46 months.

Univariate analysis results

In the study, among the 228 children diagnosed with Kawasaki disease, 40 cases (17.54%) experienced the complication of CAL. The univariate analysis showed that several factors, including sex, WBC, PLT, CKMB, CRP, ALB, and neutrophil, significantly influenced the development of CAL in Kawasaki disease (Table 1).

Table 1 Univariable analysis of risk factors for KD complicating CAL

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>CAL (n = 40)</th>
<th>Non-CAL (n = 188)</th>
<th>t/x²</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (n, %)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>8 (20%)</td>
<td>79 (42.02%)</td>
<td>5.88</td>
<td>0.02</td>
</tr>
<tr>
<td>Female</td>
<td>32 (80%)</td>
<td>109 (57.98%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity (n, %)</td>
<td></td>
<td></td>
<td>2.22</td>
<td>0.14</td>
</tr>
<tr>
<td>Ethnic minorities</td>
<td>8 (20%)</td>
<td>19 (10.11%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Han nationality</td>
<td>32 (80%)</td>
<td>169 (89.89%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

KD, Kawasaki disease; CAL, coronary artery lesions.

Submit a manuscript: https://www.tmjournals.com/mdm
Table 1 Univariable analysis of risk factors for KD complicating CAL (Continued)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>CAL (n = 40)</th>
<th>Non-CAL (n = 188)</th>
<th>χ²</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Months (x̅ ± s, months)</td>
<td>36.58 ± 23.95</td>
<td>43.49 ± 29.24</td>
<td>-1.59</td>
<td>0.12</td>
</tr>
<tr>
<td>WBC (x̅ ± s, 10³/L)</td>
<td>15.43 ± 7.14</td>
<td>11.36 ± 4.03</td>
<td>3.49</td>
<td>0.001</td>
</tr>
<tr>
<td>PLT (x̅ ± s, 10³/L)</td>
<td>489.15 ± 196.02</td>
<td>401.82 ± 172.99</td>
<td>2.61</td>
<td>0.01</td>
</tr>
<tr>
<td>RBC (x̅ ± s, 10³/L)</td>
<td>4.01 ± 0.52</td>
<td>4.15 ± 0.45</td>
<td>-1.64</td>
<td>0.11</td>
</tr>
<tr>
<td>Hb (x̅ ± s, g/L)</td>
<td>105.78 ± 12.93</td>
<td>109.74 ± 14.91</td>
<td>-1.71</td>
<td>0.09</td>
</tr>
<tr>
<td>Lymphocyte (x̅ ± s, 10³/L)</td>
<td>4.3 ± 2.46</td>
<td>3.67 ± 1.82</td>
<td>1.53</td>
<td>0.13</td>
</tr>
<tr>
<td>Neutrophil (x̅ ± s, 10³/L)</td>
<td>9.58 ± 6.93</td>
<td>6.32 ± 2.65</td>
<td>2.93</td>
<td>0.01</td>
</tr>
<tr>
<td>Monocyte (x̅ ± s, 10³/L)</td>
<td>0.99 ± 0.83</td>
<td>0.88 ± 0.84</td>
<td>0.72</td>
<td>0.48</td>
</tr>
<tr>
<td>CRP (x̅ ± s, mg/L)</td>
<td>89.56 ± 66.72</td>
<td>65.53 ± 55.68</td>
<td>2.13</td>
<td>0.04</td>
</tr>
<tr>
<td>DBIL (x̅ ± s, umol/L)</td>
<td>8.96 ± 9.59</td>
<td>10.62 ± 9.07</td>
<td>-1.00</td>
<td>0.32</td>
</tr>
<tr>
<td>ALB (x̅ ± s, g/L)</td>
<td>31.02 ± 5.10</td>
<td>33.75 ± 4.94</td>
<td>-3.09</td>
<td>0.003</td>
</tr>
<tr>
<td>ALT (x̅ ± s, U/L)</td>
<td>46.98 ± 59.93</td>
<td>55.86 ± 82.68</td>
<td>-0.79</td>
<td>0.43</td>
</tr>
<tr>
<td>AST (x̅ ± s, U/L)</td>
<td>38.59 ± 38.54</td>
<td>46.24 ± 60.87</td>
<td>-1.01</td>
<td>0.31</td>
</tr>
<tr>
<td>CKMB (x̅ ± s, U/L)</td>
<td>15.22 ± 6.91</td>
<td>19.14 ± 10.57</td>
<td>-2.93</td>
<td>0.004</td>
</tr>
<tr>
<td>PT (x̅ ± s, s)</td>
<td>13.12 ± 4.91</td>
<td>12.21 ± 2.59</td>
<td>1.14</td>
<td>0.26</td>
</tr>
<tr>
<td>APTT (x̅ ± s, s)</td>
<td>30.91 ± 4.67</td>
<td>30.43 ± 5.77</td>
<td>0.57</td>
<td>0.57</td>
</tr>
</tbody>
</table>

KD, Kawasaki disease; CAL, coronary artery lesions; WBC, white blood cell count; PLT, platelet; RBC, red blood cell count; Hb, hemoglobin concentration; CRP, C-reactive protein; DBIL, direct bilirubin; ALB, albumin; ALT, alanine transaminase; AST, aspartate transaminase; CKMB, creatine kinase isoenzyme MB; PT, prothrombin time; APTT, activated partial thromboplastin time.

Machine learning results

Regarding the 18 predictive factors included in the study, the LASSO algorithm identified 9 predictive variables: WBC, PLT, neutrophil, ALB, CKMB, months, ethnicity, sex and prothrombin time (Figure 1A, 1B). Before proceeding with the random forest algorithm, we used 5-fold cross-validation to find the optimal K value as 10, with an root mean square error (RMSE) of 0.308. Then we identified 8 predictive factors with an increase in node purity value above 1.5: CKMB, alanine transaminase, red blood cell count, hemoglobin concentration, activated partial thromboplastin time, ALB, WBC, and neutrophil (Figure 1C). The model's R² value of 0.826 indicates a good level of fit. By taking the intersection of these results, a final set of 4 predictive factors (WBC, ALB, CKMB and neutrophil) was determined for further analysis. These factors were included in a logistic regression model to evaluate their predictive value, and a forest plot was generated based on the regression model.

Model evaluation

We evaluated the predictive value of each factor by calculating the AUC. The AUC values for WBC (AUC = 0.686), ALB (AUC = 0.65), CKMB (AUC = 0.642), and neutrophil (AUC = 0.656) demonstrated their individual predictive abilities (Figure 2A). The model created by incorporating these four predictive factors together yielded an AUC of 0.743 (Figure 2B), indicating a good overall predictive performance. The sensitivity, specificity, accuracy, positive predictive value, and negative predictive value of WBC, ALB, CKMB, neutrophil, and their combination are presented in detail in the following Table 2. The combined metrics were superior to those of individual factors. Calibration curves showed satisfactory consistency for the risk prediction model of CAL (Figure 2C). The DCA curve further confirmed the accuracy of the model (Figure 2D).

Discussion

KD is a self-limiting febrile illness that typically affects children under the age of 5, particularly those of Asian descent. Common symptoms of KD include high fever, rash, inflammation of the oral mucosa, conjunctivitis, and swollen lymph nodes [2]. It primarily affects the cardiovascular system, particularly the coronary arteries, leading to arteritis and vasculitis. While the exact cause and mechanism of KD are still unclear, it has become a common cause of acquired heart disease in children [8–10]. Therefore, investigating the clinical characteristics and risk factors for CAL in KD patients is essential for more effective treatment of high-risk patients and improving their prognosis.
Figure 1 Using multiple machine learning methods to select the best predictive factors. (A) The LASSO coefficient profile was utilized to analyze candidate predictive factors. (B) Optimal penalty coefficients were determined through LASSO regression. (C) Random forest feature selection was performed to obtain the selected features. (D) A risk nomogram was developed specifically for predicting Kawasaki disease with concurrent CAL. CAL, coronary artery lesions; WBC, white blood cell count; ALB, albumin; APTT, activated partial thromboplastin time; Hb, hemoglobin concentration; RBC, red blood cell count; ALT, alanine transaminase; CKMB, creatine kinase isoenzyme MB; PT, prothrombin time; PLT, platelet; DBIL, direct bilirubin; AST, aspartate transaminase; CRP, C-reactive protein; LASSO, Least Absolute Shrinkage and Selection Operator.
In this study, we collected clinical data from 228 KD patients retrospectively. Univariate analysis revealed a significant correlation between CAL occurrence and several factors, including sex, WBC, PLT, CKMB, CRP, ALB, and neutrophil. Using LASSO and random forest algorithms, we selected WBC, CKMB, ALB, and neutrophil count as the final predictive factors among 18 initial predictors. Based on these four factors, we developed nomograms, and their discriminative ability and accuracy were validated using the AUC, calibration curves, and DCA curve.

In patients with KD, the presence of CAL is a severe and important
complication. Research has shown that in the early stages of KD, it primarily presents as systemic microvasculitis. After approximately two weeks, the disease progresses to involve inflammation of the arterial intima and periarterial tissues in the branches of the aorta, resulting in damage to the elastic layer within the blood vessels and a reduction in their structural integrity. The coronary arteries, in particular, experience significant inflammatory injury, including the rupture of elastic fibers in the arterial intima. The severity of CAL can vary among patients, with mild cases potentially resolving spontaneously, while severe cases can lead to inadequate blood supply, abnormal dilation of the coronary arteries, and even blood clot formation. In severe situations, CAL can cause myocardial ischemia, myocardial infarction, and other cardiac issues. It is important to note that most deaths in KD patients are attributed to CAL. Therefore, early identification and monitoring of CAL are crucial for KD patients. Research by Gong et al. has indicated that gamma-glutamyl transferase, the neutrophil-to-lymphocyte ratio, and the size of aneurysms are independent risk factors for persistent CAL [11]. Another study has demonstrated that the neutrophil-to-lymphocyte ratio is a reliable predictor of CAL [12]. Additionally, markers such as CRP, red blood cell distribution width, tumor necrosis factor-alpha, and IgA concentration have been found to be useful in predicting CAL [13-18]. Furthermore, the incidence of CAL is higher in atypical types of KD, such as incomplete KD and KD with no response to intravenous immunoglobulin, compared to typical KD [18].

KD is characterized by changes in white blood cell count. In the early stages of the disease, the count may be normal or slightly elevated. However, as inflammation progresses, the count tends to increase and remain high. An elevated neutrophil count is commonly observed in KD. Neutrophils, the most prevalent white blood cell type, make up about 60–70% of the total count. They play a crucial role in maintaining immune system function, protecting health, and combating pathogens, infection, and inflammation. The neutrophil-to-lymphocyte ratio (NLR) has shown promise as a biomarker for predicting CAL in KD [12]. Research by Chidambaram et al. indicates that an NLR ≥ 2.08 between the fourth and sixth day of fever can help identify the risk of CAL. However, current research on NLR in KD has limitations, such as variations in population and ethnicity, and small sample sizes in single-center retrospective studies. Consequently, larger-scale multicenter studies are needed to determine the optimal predictive value of NLR for coronary artery dilation in KD. Recent research has discovered that in patients with KD, there is an increased formation of neutrophil-generated neutrophil extracellular traps [19]. This heightened formation leads to a significant increase in cellular activity in peripheral blood mononuclear cells, inhibits cell apoptosis, and boosts the production of pro-inflammatory cytokines as well as the activation of NF-kB [19, 20]. Additionally, the stimulation of neutrophil extracellular traps also results in elevated expression of vascular endothelial growth factor A and hypoxia-inducible factor-1α, thereby further intensifying vascular inflammation in KD [21].

ALB is one of the most prevalent proteins in the blood. Inflammatory reactions caused by KD can result in an elevation of white blood cells and inflammatory mediators, leading to ongoing depletion of albumin within the body [9]. Furthermore, vascular injury and vasculitis can heighten the permeability of blood vessel walls, causing albumin to leak from the bloodstream into surrounding tissues and subsequently reducing the concentration of albumin in the blood. A decreased level of albumin may indicate more severe vascular damage and an increased likelihood of developing CAL. There are several studies suggesting that lower levels of ALB are an independent risk factor for the occurrence of CAL in patients with KD [5, 9, 10, 14]. In KD, the occurrence of inflammation in the coronary arteries can potentially lead to a series of myocardial injury events. However, direct myocardial damage is relatively uncommon [10]. As a result, in cases of Kawasaki disease, elevated blood elevation or even normal levels of CKMB are frequently observed. Currently, there is a lack of a well-established scoring system for predicting CAL in Kawasaki disease. Our study aims to develop a CAL prediction model specifically for Kawasaki disease patients at our hospital, but there are limitations that need to be considered. The main limitations include the retrospective study design and the limited number of patients. In order to validate the predictive effectiveness of the nomogram, larger-scale prospective studies are necessary. Additionally, retrospective studies inevitably have selection biases. Lastly, although we innovatively discussed the impact of ethnicity on CAL prediction, we only categorized it into two broad groups, the Han ethnicity and other ethnic minorities, without further subdivision.

References


