Discover Oncology

Research

Development of a nomogram for predicting the risk of carcinoma in chronic atrophic gastritis

Jia-Yi Zhang[1,2] · Ding Li[2] · Guo-Jie Hu[2]

Received: 30 November 2024 / Accepted: 21 April 2025
Published online: 08 May 2025

Abstract

Objective To construct a machine learning (ML) model to predict the progression of chronic atrophic gastritis (CAG) to gastric cancer (GC), given its precancerous significance.

Methods Using medical records from the Affiliated Hospital of Qingdao University, common laboratory indicators were extracted. LASSO regression identified 10 core risk factors, which were further analyzed using binary logistic regression to develop a nomogram model in R. The model's performance was evaluated using receiver operating characteristic (ROC) curves, the concordance index (C-index), calibration curves, and decision curve analysis (DCA).

Results The model showed excellent performance, with a C-index of 0.887. The key factors included sex, coagulation, blood cell indexes, and blood lipid levels. The ROC areas were 0.892 (quantitative) and 0.853 (qualitative), confirming model reliability.

Conclusion A new nomogram model for assessing GC risk in CAG patients was successfully developed. However, due to data collection and time limitations, future studies should expand the sample size, perfect the validation process, and optimize the model to achieve more accurate risk prediction.

Keywords Chronic atrophic gastritis · Gastric cancer · Machine learning · Predictive model · Nomogram

Abbreviations

CAG	Chronic atrophic gastritis
GC	Gastric cancer
ML	Machine learning
LASSO	Least absolute shrinkage and selection operator
ROC	Receiver operating characteristic
AUC	Area under the curve
DCA	Decision curve analysis
CI	Confidence interval
OR	Odds ratios
C-index	The concordance index
MLE	Maximum Likelihood Estimation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12672-025-02453-y.

✉ Guo-Jie Hu, huguojie2003@163.com; Jia-Yi Zhang, zjiayi1020@163.com; Ding Li, 907325791@qq.com | [1]Institute of Integrated Medicine, Qingdao Medical College of Qingdao University, Qingdao University, Qingdao, Shandong, China. [2]Department of Traditional Chinese Medicine, The Affiliated Hospital of Qingdao University, Qingdao, Shandong, China.

1 Introduction

Gastric cancer (GC) is a malignant tumor that poses a significant threat to human health globally, with morbidity and mortality rates ranking among the highest of all cancer types [1, 2]. Chronic atrophic gastritis (CAG) is recognized as a crucial precursor lesion for GC [3]. The progression from CAG to GC follows a specific pathological pathway. Correa et al. [4] proposed that GC develops through a multistep process, beginning with the transformation of normal gastric mucosa into chronic nonatrophic gastritis, which subsequently progresses to CAG, followed by intestinal metaplasia, dysplasia, and ultimately, GC. This theory provides a vital framework for understanding the pathogenesis of GC and underscores the importance of accurately assessing cancer risk at the CAG stage.

Currently, the clinical assessment of GC risk in CAG patients primarily depends on regular endoscopic examinations [5] and tumor marker detection [6]. However, endoscopy is an invasive procedure with low patient acceptance, and frequent examinations can cause significant physical discomfort and incur high economic costs. Although tumor marker detection is relatively convenient, the specificity and sensitivity of certain markers still require enhancement. Consequently, accurate assessment of the risk of GC relying solely on tumor markers is challenging. Additionally, many primary medical institutions face limitations in equipment and technology, making it difficult to routinely conduct high-quality endoscopy and precise tumor marker detection. Thus, existing assessment methods may have practical limitations and fail to meet the clinical demand for efficient and accurate evaluation of GC risk in CAG patients.

In the medical field, machine learning (ML) technology has experienced rapid advancements in recent years and has demonstrated significant potential in disease risk assessment, diagnosis, and prediction [7]. ML algorithms are capable of thoroughly mining and analyzing extensive amounts of complex medical data, uncovering hidden patterns and rules, and establishing accurate predictive models [8]. The advantages of ML include its ability to handle nonlinear relationships, automatically screen for important features, mitigate the influence of human factors, and provide a novel perspective and powerful tools for medical research and clinical practice [9]. For example, previous studies have successfully utilized ML to analyze key genes that predict the progression of CAG to GC and to identify signaling pathways associated with these genes [10]. This finding not only enhances our understanding of the carcinogenic mechanisms underlying CAG but also offers clinicians new treatment strategies.

Given this context, the present study aimed to develop a cost-effective, efficient, and accurate model for assessing the risk of GC in patients with CAG. By extracting potential information from clinical data, ML algorithms were employed to identify key risk factors among common biological indicators, thereby constructing a reliable assessment model to aid clinicians in implementing stratified management of CAG patients and facilitating early prevention and treatment of GC through precise interventions.

2 Materials and methods

2.1 Case collection

This study analyzed the medical records of patients diagnosed with CAG and GC who were admitted to the Affiliated Hospital of Qingdao University between January 1, 2016, and December 31, 2022. The hospital offered robust data support for this research, leveraging its extensive clinical case resources and advanced medical technology.

The inclusion criteria for patients with CAG were as follows: the endoscopic pathological examination reports of the included patients met the internationally recognized diagnostic standards for CAG and clearly demonstrated atrophy of the intrinsic glands of the gastric mucosa, with or without intestinal epithelialization. The presence of abnormal growth, dysplasia, and inflammatory cell infiltration aligned with the pathological characteristics of CAG [11]. Additionally, the diagnosis was independently reviewed and confirmed by at least two senior pathologists to ensure accuracy.

Patients with chronic nonatrophic gastritis, whose gastric mucosa primarily exhibited chronic inflammatory cell infiltration predominantly composed of lymphocytes and plasma cells, without inherent glandular atrophy, were excluded [11]. Patients with malignant gastric tumors, including pathologically confirmed gastric adenocarcinoma, gastric lymphoma, and other types of gastric malignancies, were also excluded. Furthermore, patients with systemic diseases that may affect gastric mucosal lesions, such as severe autoimmune diseases and those receiving long-term immunosuppressive treatment, were excluded to ensure the homogeneity of the research subjects and minimize the interference of extraneous factors with the research outcomes.

The inclusion criteria for GC patients were as follows: clinically diagnosed with malignant gastric tumors based on the criteria outlined in the latest version of the World Health Organization Classification of Digestive System Tumors [12], along with gastroscopic manifestations, histopathological examination, and immunohistochemistry. Various methods, including chemical analyses, were utilized to confirm the characteristics of the tumor cells, such as atypia and infiltrative growth. Immunohistochemical testing was employed to further delineate the histological type and degree of differentiation of the tumor, thereby providing a more robust basis for accurate diagnosis. Additionally, a detailed medical history, comprehensive physical examination, and necessary imaging studies—such as whole-body PET—CT and abdominal contrast-enhanced CT—were conducted to rule out the possibility of malignant tumors in other regions of the patient metastasizing to the stomach, thereby ensuring the integrity of the GC case group and minimizing interference from extraneous factors in the research results.

All pathological data were duly approved by the Department of Pathology and the Department of Gastroenterology.

2.2 Risk factor screening

Among the collected clinical case data, we selected sex, age, and 20 common laboratory indicators as relevant factor variables. These variables underwent a rigorous preprocessing process before being utilized to build the model. First, all the indicators were quantified, and various forms of data were converted into a unified numerical format suitable for mathematical analysis, ensuring consistency and comparability. Simultaneously, techniques such as data interpolation and the deletion of samples with missing values were employed to avoid data gaps, thereby ensuring data integrity and establishing a solid foundation for subsequent accurate data analysis.

This study utilized SPSS version 26.0 and R software (version 4.3.0) for data analysis. In R, the "glmnet" package was employed to conduct least absolute shrinkage and selection operator (LASSO) [13] regression analysis on the 22 indicators (including sex, age, and the 20 laboratory indicators). During the LASSO regression, the parameter was set to 999. This parameter setting was validated through multiple preexperiments and theoretical considerations, effectively screening the most representative variables from a large dataset while maintaining model stability. Through the application of the LASSO regression algorithm, the optimal risk factors were identified from the organized data. During the screening process, we focused on variables with nonzero coefficients in the LASSO regression, which are considered risk factors closely associated with the risk of GC in CAG patients and served as core variables for subsequent model construction.

2.3 Model construction

This study incorporated ten independent variables, including sex, coagulation function indicators, and hematological parameters, to construct a predictive model using the confirmed status of GC as a binary dependent variable (1 = GC group, 0 = CAG group). A binary logistic regression model was employed, with regression coefficients for each variable estimated using the Maximum Likelihood Estimation (MLE) method to obtain the predicted probability of GC occurrence. During the model parameter estimation process, model calibration was assessed using the Hosmer–Lemeshow test ($\alpha = 0.05$), and the Odds Ratio (OR) along with its 95% confidence interval were calculated. The diagnostic efficacy was validated using the Receiver Operating Characteristic curve (ROC), based on the predicted probabilities from the model. The analytical methods included: (1) treating predicted probabilities as continuous test variables and actual diagnostic results as binary state variables (positive threshold = 1); (2) calculating the Area Under the Curve (AUC) and its 95% Confidence Interval (95% CI) using non-parametric methods; (3) interpreting diagnostic value according to internationally recognized standards: an AUC of 0.5 indicates no discriminative ability, 0.7–0.8 suggests moderate diagnostic efficacy, 0.8–0.9 represents good discriminative power, and > 0.9 denotes excellent diagnostic value. The nomogram model integrates key risk factors to visually demonstrate the quantitative relationship between these factors and the risk of GC onset in a graphical format. In constructing the nomogram, the "rms" and "rmda" packages in R software are utilized to calculate the model's concordance index (C-index) [14], calibration curve [15], and decision curve analysis (DCA) [16]. By visualizing the selected key factors and their corresponding regression coefficients, the model's performance is comprehensively evaluated and validated to ensure its accuracy and reliability [17].

2.4 Statistical analysis

In this study, SPSS 26.0 was used for statistical analysis. A *p-value* < 0.05 (two-tailed) was considered statistically significant.

During the analysis, some variables with $p < 0.05$ in the LASSO regression showed $p > 0.05$ in the logistic regression. This difference may be due to the different principles and application scenarios of the two regression methods. The LASSO regression can optimize the model while screening variables, which may cause some variables that are not significant in the logistic regression to be significant. Therefore, for the variables with $p > 0.05$ in the logistic regression, although they did not reach the significance standard of 0.05, they still have reference value in medical research. These findings may imply a weak association with the research results. In this study, these variables were selected to construct the model.

3 Results

3.1 Basic characteristics of the data

According to the established inclusion and exclusion criteria, a total of 350 patients were enrolled in this study, comprising 150 patients diagnosed with CAG and 200 patients with GC. Among the participants, 207 were male, and 143 were female. A statistical analysis was conducted on 20 common laboratory indicators, with the *chi-square values* and *p-values* for each indicator displayed in Table 1.

Following the LASSO regression analysis conducted on the aforementioned data, we determined that 10 out of the 22 factor variables exhibited nonzero coefficients in the established model (Fig. 1A and B). The identified risk factors include sex, prothrombin time, fibrinogen, D-dimer, the TT ratio, the mean corpuscular volume, hemoglobin, platelets, high-density lipoprotein, and triglycerides.

3.2 Model structure

On the basis of the ten factor variables screened above, we successfully constructed a binary logistic regression model. Using SPSS 26.0 software for in-depth analysis, we identified the risk factors associated with the progression of CAG to GC, and the specific data are presented in Table 2. The C-index of the model was calculated to be 0.895 (95% CI 0.863–0.927), providing strong evidence of the model's predictive ability. Furthermore, we conducted a detailed analysis of each feature, which included the calculation of its 95% confidence interval (CI), the OR within the 95% CI, and the corresponding *p-value*.

On the basis of the predictor variables mentioned above, we constructed a nomogram to assess the risk of cancer in patients with CAG, as illustrated in Fig. 2. When using this nomogram, for instance, If the patient is male, mark the corresponding sex on the sex line. Subsequently, repeat this operation for the other identified variables. After marking all the relevant points, draw perpendicular lines to each variable line until they intersect with the "point" line. The values corresponding to each intersection point on the "point" scale are the predicted scores. Summarize these scores to obtain the "total point", and mark this point on the scale. Then, draw a perpendicular line from this point to the scale until it intersects with the "risk of gc" scale. The final intersection point indicates the predicted probability of the patient developing GC.

To further validate the performance of the nomogram, we analyzed it using SPSS 26.0 software. The results indicated that the AUCs of the qualitative and quantitative ROC curves were 0.853 and 0.892 (Fig. 3), respectively. These findings suggest that the nomogram has a strong ability to predict the likelihood of cancer development in patients with CAG.

The calibration curve of the nomogram (Fig. 4A) exhibited a positive tendency in its calibration performance for predicting the probability of GC development. In the intermediate-probability range, the bias-corrected solid line closely aligns with the ideal diagonal line. This alignment clearly indicates that the model is capable of predicting the probability of GC occurrence with relatively high precision within this range.

Notably, in the low-probability and high-probability regions at the two ends, the solid line deviates to some extent from the ideal curve. This deviation signifies that the model's accuracy declines when dealing with the prediction of extremely low or high probabilities of GC. This finding also implies that there could be certain disparities in the prediction results, which necessitates careful consideration in practical clinical applications. These findings underscore the

Table 1 Included Risk Factors and Their Corresponding *Chi-Square Values* and *P-values*

Factor	Variables	CAG Group	GC Group	χ^2	*P*
		n = 150	n = 200		
Sex	Male	66 (44.0%)	141 (70.5%)	24.910	0.000
	Female	84 (56.0%)	59 (29.5%)		
year	< 60	51 (34.0%)	56 (28.0%)	1.454	0.228
	≥ 60	99 (66.0%)	144 (72.0%)		
Prothrombin time(sec)	10–14	106 (70.7%)	118 (59.0%)	7.656	0.022
	< 10	44 (29.3%)	77 (38.5%)		
	> 14	0 (0.0%)	5 (2.5%)		
Fibrinogen (g/L)	2–4	140 (93.3%)	147 (73.5%)	33.905	0.000
	< 2	6 (4.0%)	3 (1.5%)		
	> 4	4 (2.7%)	50 (25.0%)		
D-dimer (ng/mL)	≤ 500	139 (92.7%)	131 (65.5%)	35.876	0.000
	> 500	11 (7.3%)	69 (34.5%)		
TT ratio	0.93–1.53	148 (98.7%)	188 (94.0%)	11.861	0.003
	< 0.93	0 (0.0%)	12 (6.0%)		
	> 1.53	2 (1.3%)	0 (0.0%)		
White blood cell count (10^9/L)	3.5–9.5	136 (90.7%)	168 (84.0%)	5.226	0.073
	< 3.5	9 (6.0%)	13 (6.5%)		
	> 9.5	5 (3.3%)	19(9.5%)		
Red blood cell count (10^9/L)	4.3–5.8	113 (75.3%)	95 (47.5%)	32.069	0.000
	< 4.3	35 (23.3%)	105 (52.5%)		
	> 5.8	2 (1.3%)	0 (0.0%)		
Mean corpuscular volume (fL)	82–100	146 (97.3%)	148 (74.0%)	35.288	0.000
	< 82	2 (1.3%)	42 (21.0%)		
	> 100	2 (1.3%)	10 (5.0%)		
Lymphocyte count (10^9/L)	1.1–3.2	136 (90.7%)	156 (78.0%)	12.482	0.002
	< 1.1	10 (6.7%)	40 (20.0%)		
	> 3.2	4 (2.7%)	4 (2.0%)		
Hemoglobin (g/L)	130–175	114 (76.0%)	88 (44.0%)	35.965	0.000
	< 130	36 (24.0%)	112 (56.0%)		
	> 175	0 (0.0%)	0 (0.0%)		
Platelet (10^9/L)	125–350	143 (95.3%)	161(80.5%)	21.159	0.000
	< 125	5 (3.3%)	7 (3.5%)		
	> 350	2 (1.3%)	32 (16.0%)		
Neutrophil count (10^9/L)	1.8–6.3	127 (84.7%)	158 (79.0%)	16.581	0.000
	< 1.8	21 (14.0%)	17 (8.5%)		
	> 6.3	2 (1.3%)	25 (12.5%)		
Cholesterol (mmol/L)	2.32–5.62	108 (72.0%)	153 (76.5%)	1.758	0.415
	< 2.32	1 (0.7%)	3 (1.5%)		
	> 5.62	41 (27.3%)	44(22.0%)		
Low-density lipoprotein (mmol/L)	< 3.37	89 (59.3%)	100 (50.0%)	73.457	0.000
	≥ 3.37	61 (40.7%)	100 (50.0%)		
High-density lipoprotein (mmol/L)	0.9–2	127 (84.7%)	182 (91.0%)	16.644	0.000
	< 0.9	6 (4.0%)	15 (7.5%)		
	> 2	17 (11.3%)	3 (1.5%)		
Triglycerides (mmol/L)	0.3–1.92	129 (86.0%)	186 (93.0%)	38.967	0.000
	< 0.3	0 (0.0%)	14 (7.0%)		
	> 1.92	21 (14.0%)	0 (0.0%)		
Aspartate aminotransfera-se (U/L)	15–40	131 (87.3%)	157 (78.5%)	7.488	0.024
	< 15	17 (11.3%)	29 (14.5%)		
	> 40	2 (1.3%)	14 (7.0%)		

Table 1 (continued)

Factor	Variables	CAG Group	GC Group	χ^2	P
		n = 150	n = 200		
Alanine aminotransfera-se (U/L)	9–50	142 (94.7%)	172 (86.0%)	7.566	0.023
	< 9	6 (4.0%)	16 (8.0%)		
	> 50	2 (1.3%)	12 (6.0%)		
Potassium (mmol/L)	3.5–5.3	146 (97.3%)	193(96.5%)	3.589	0.166
	< 3.5	4 (2.7%)	3(1.5%)		
	> 5.3	0 (0.0%)	4(2.0%)		
uric acid (μmol/L)	89.2–416	133 (88.7%)	170 (85.0%)	2.107	0.349
	< 89.2	0 (0.0%)	2 (1.0%)		
	> 416	17 (11.3%)	28 (14.0%)		
Glucose (mmol/L)	3.9–6.16	134 (89.3%)	161 (80.5%)	6.035	0.049
	< 3.9	3 (2.0%)	13 (6.5%)		
	> 6.16	13 (8.7%)	26 (13.0%)		

CAG chronic atrophic gastritis, *GC* gastric cancer

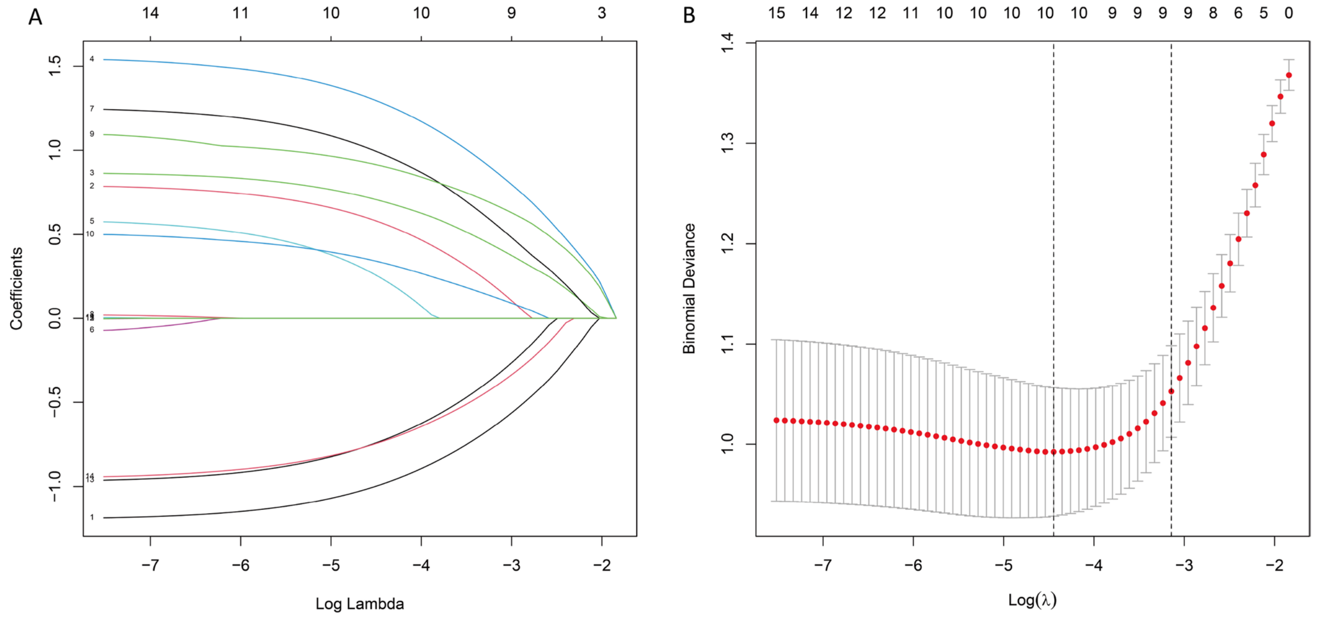

Fig. 1 The coefficient path diagram of risk factor variables obtained through LASSO regression analysis illustrates two key aspects. **A** Coefficient trajectories showing variable weight changes with increasing regularization intensity Log(λ). **B** Optimal λ selection yielding 10 non-zero coefficient predictors. *LASSO* Least Absolute Shrinkage and Selection Operator

importance of understanding the performance characteristics of the nomogram across different probability ranges to optimize its clinical utility.

The DCA of the nomogram depicted in Fig. 4B provides a clear guide for evaluating the clinical utility of the model. When the threshold probability falls within the critical range spanning from 0.04 to 0.96, the red dotted line, which represents the decision curve of the GC prediction nomogram, resides above the gray dotted line (denoted as "All") and the black dotted line (denoted as "None"), with these two latter lines signifying extreme intervention strategies. Conversely, outside this threshold, the red dotted line falls below the other two lines, suggesting a diminished decision-making advantage.

The threshold probability, for both patients and doctors, pertains to the probability value derived from the nomogram prediction. This value assumes a pivotal role in clinical decision-making processes. For patients, when the probability of GC occurrence, as predicted by the nomogram, is within the 0.04–0.96 threshold range, patients are more likely to be inclined toward undergoing further invasive examinations, such as gastroscopic biopsy, or opting to initiate a preventive

Table 2 The regression coefficients, Odds ratio, 95% CIs, and *p-values* for each risk factor

Variable	Prediction model		
	β	Odds ratio(95% CI)	*p-value*
Sex	– 1.271	0.281 (0.155–0.509)	0.000
Prothrombin time	– 0.314	0.731 (0.576–0.927)	0.010
Fibrinogen	0.721	2.056 (1.242–3.406)	0.005
D-dimer	0.000	1.000 (1.000–1.001)	0.027
TT ratio	– 6.004	0.002 (0.000–0.074)	0.001
Mean corpuscular volume	– 0.035	0.966 (0.916–1.017)	0.190
Hemoglobin	– 0.037	0.964 (0.948–0.980)	0.000
Platelet	0.004	1.004 (0.999–1.008)	0.118
High-density lipoprotein	– 1.169	0.311 (0.136–0.709)	0.005
Triglycerides	– 0.836	0.434 (0.252–0.746)	0.003

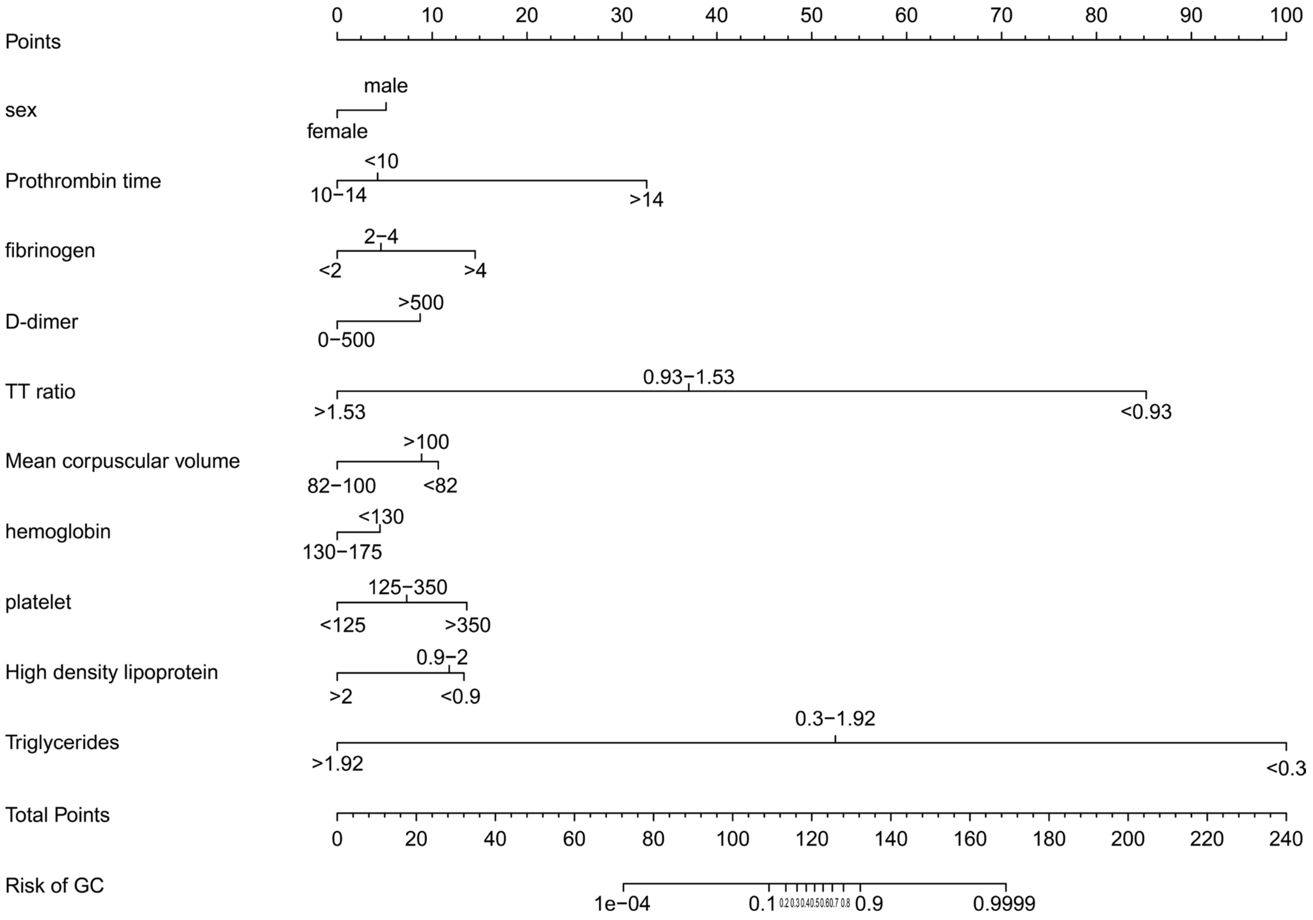

Fig. 2 Prediction model for GC risk assessment in CAG patients. *CAG* Chronic atrophic gastritis, *GC* Gastric Cancer

treatment plan. From the perspective of the doctors, this threshold probability serves as a crucial reference for them to determine whether more proactive clinical intervention measures should be considered. For example, when the predicted probability exceeds 0.04, doctors might recommend that patients shorten their follow-up cycle or adopt more aggressive treatment measures for CAG to mitigate the risk of cancer development.

Within this threshold range, in contrast to the indiscriminate application of a unified "all-intervention" or "no-intervention" approach for all patients, leveraging the nomogram model to facilitate decision-making is capable of yielding more substantial net benefits in clinical practice. It effectively assists doctors in formulating personalized and precise diagnosis and treatment plans for patients. However, once the threshold probability exceeds this range,

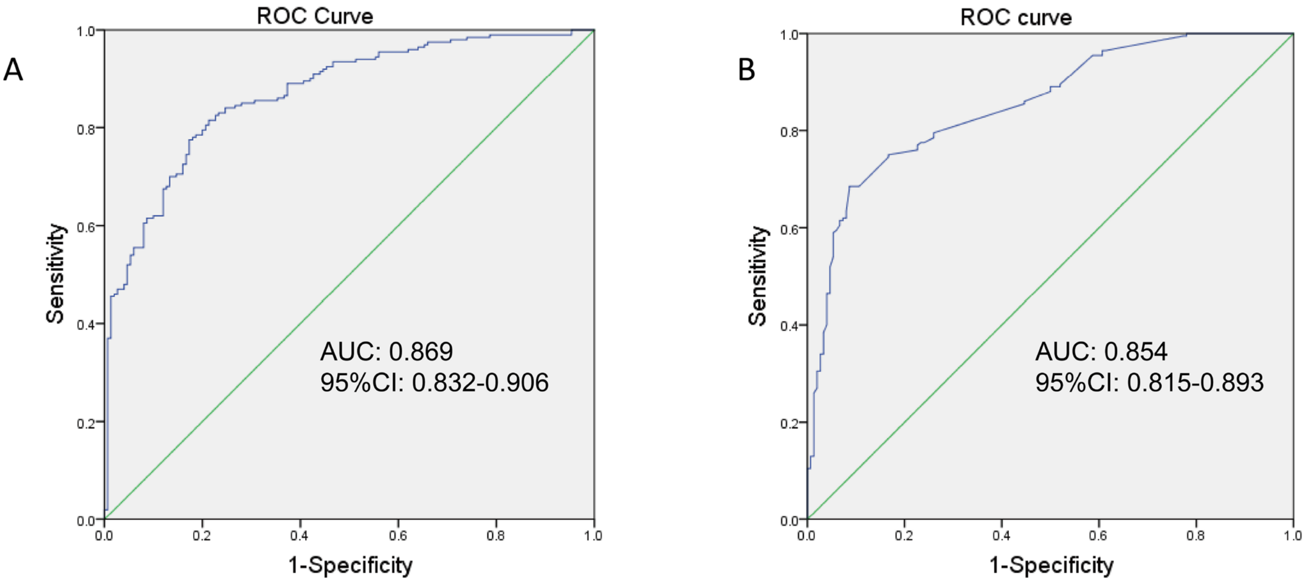

Fig. 3 The ROC curve pertaining to the nomogram model was analyzed. **A** The quantitative manifestation of the ROC curve for the nomogram model is presented, wherein the AUC attains a value of 0.869, accompanied by a 95% CI spanning from 0.832 to 0.906. **B** Depicting the qualitative aspect of the ROC curve for the same nomogram model, revealing an AUC of 0.854 and a 95% CI lying between 0.815 and 0.893. The ROC curve: the Receiver Operating Characteristic curve, *the AUC* the Area Under the Curve, *95% CI* 95% confidence interval

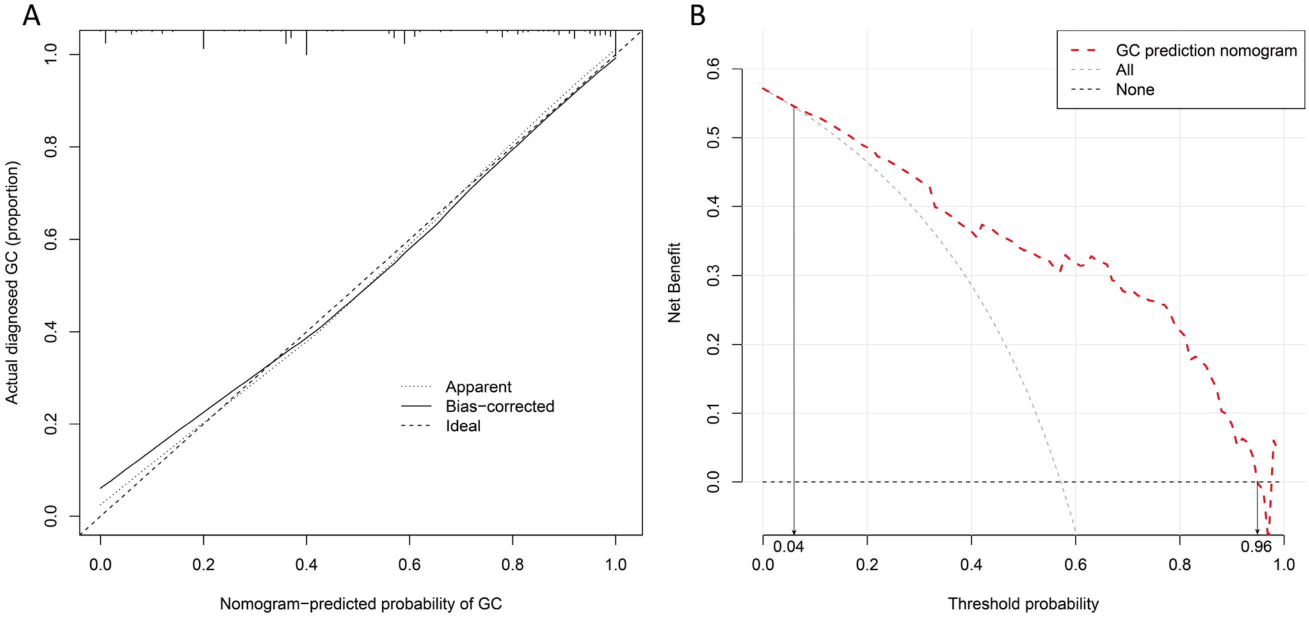

Fig. 4 Calibration curve and decision curve for the nomogram. **A** Nomogram Calibration Curve. This curve comprises three distinct lines: the dotted line (Apparent) represents the original calibration curve; the solid line (Bias-corrected) reflects the curve after bias correction and closely aligns with the ideal diagonal line (Ideal) in the intermediate probability region. **B** Nomogram DCA. Among the three curves, the red dotted line (GC prediction nomogram) illustrates the decision curve of the nomogram. When the threshold probability is between 0.04 and 0.96, this line exceeds the gray dotted line (All), which signifies "full intervention", and the black dotted line (None), representing "no intervention". *GC* gastric cancer, *DCA* decision curve analysis

the red dotted line declines precipitously and drops below the gray or black dotted line. This indicates that the model is ill suited to serve as a primary decision-making tool for those circumstances. Otherwise, it could lead to the irrational allocation of diagnostic and treatment resources or the forfeiture of the optimal intervention opportunity.

By integrating the information gleaned from the two figures, the nomogram model demonstrated significant value in the domain of GC prediction. Its outstanding performance in intermediate-probability calibration and its remarkable decision-making advantages within a specific threshold range offers robust support for the early screening, disease assessment, and precise intervention of GC. Although the model exhibits certain limitations in extreme-probability prediction and decision-making beyond the threshold, through the rational demarcation of its scope of application, its advantages could still be fully harnessed to contribute to enhancing the level of GC diagnosis and treatment.

4 Discussion

The early and accurate assessment of carcinogenesis in CAG has long been a critical and pressing issue in the realm of clinical practice. However, conventional methods are frequently encumbered with certain limitations. In this study, we endeavored to develop a novel risk assessment model by integrating ML algorithms with multivariate statistical analysis techniques [18].

Within this model, we meticulously screened out a comprehensive set of key risk factors. These factors include sex, prothrombin time, fibrinogen, D-dimer, the TT ratio, the mean corpuscular volume, hemoglobin, platelets, high-density lipoprotein, and triglycerides. Notably, the indicators predicted by our model have also been the subject of significant attention in prior research. This not only validates the rationality of our model's construction but also supports its potential to contribute to the existing body of knowledge in this field.

Sex plays a nonnegligible role in the carcinogenesis process of CAG. As mentioned by Xiaoyan Luan et al. [19], the incidence of GC in men is nearly twice that in women. This phenomenon may be associated with multiple factors. In terms of physiological mechanisms, estrogen secreted by women may play a certain inhibitory role in the occurrence and development of GC. Estrogen can regulate the expression of cell cycle-related proteins, thereby influencing the viability of GC cells and promoting apoptosis, which reduces the probability of transformation from CAG to GC [20]. There are differences in unhealthy lifestyle habits such as smoking and alcohol consumption between men and women, and these differences also affect the carcinogenesis process of CAG. For example, the proportion of male smokers is generally greater than that of female smokers [21], and smoking has been proven to be a risk factor for GC [22]. Harmful substances in tobacco can damage the gastric mucosa and promote *Helicobacter pylori* infection [23], further increasing the occurrence and carcinogenesis risk of CAG. Long-term heavy alcohol consumption can also irritate and damage the gastric mucosa [24], leading to inflammatory reactions in the gastric mucosa and abnormal cell hyperplasia, accelerating the transformation from CAG to GC. In contrast, women generally have fewer unhealthy lifestyle habits, which may relatively reduce their risk of CAG transformation.

Coagulation-related indicators such as prothrombin time, fibrinogen, D-dimer, and the TT ratio were incorporated into the model, suggesting that abnormalities in the coagulation system may play a pivotal role in the carcinogenesis process of CAG. Previous studies have demonstrated that tumor growth and metastasis are often accompanied by alterations in coagulation function [25, 26]. Our findings further support this view and provide new evidence for exploring the relationship between coagulation mechanisms and carcinogenesis.

Blood cell-related indicators, including the mean corpuscular volume, hemoglobin level, and platelet count, reflect the body's hematopoietic function and blood status. The association of these genes with the risk of carcinogenesis in CAG may imply that, during the disease development process, corresponding changes occur in the body's hematopoietic microenvironment and blood components. These changes may be involved in the initiation or promotion of carcinogenesis. Although some studies have focused on the associations between hematological parameters and tumors, such as abnormal changes in hematological indicators (e.g., mean corpuscular volume, hemoglobin, and platelet count) and the prognosis of patients with GC [27, 28], relevant research on the specific process of carcinogenesis in CAG is relatively scarce. In our model, these hematological parameters were screened as potential risk factors. However, owing to the current lack of research, the specific mechanisms of action and value of these parameters in the assessment of the risk of carcinogenesis in patients with CAG are not fully understood. This also highlights the importance of our exploration in this field and the challenges we face. More research efforts are needed in the future to fill this knowledge gap.

As lipid-related indicators, high-density lipoprotein and triglyceride levels emerged in the model, indicating that abnormal lipid metabolism may be closely associated with the carcinogenesis of CAG. Other studies have explored the relationship between dyslipidemia and tumorigenesis [29, 30], and our findings echo these studies. Triglycerides made a significant contribution to the nomogram model we constructed. When the triglyceride level was lower than 0.3 mmol/L, 100 points were assigned in the nomogram model. However, this model assesses the probability of GC risk on the basis of a comprehensive evaluation of multiple indicators. This 100-point score needs to be added to the scores corresponding to other indicators, such as sex, prothrombin time, and fibrinogen, to obtain the final risk prediction value. An extremely low triglyceride level reflects possible abnormal lipid metabolism in the body, which may affect the physiological function of the gastric mucosa and thus increase the risk of GC. Nevertheless, it is not the sole determinant. We are well aware of the complexity of this mechanism, and more in-depth research is needed in the future to clarify the underlying relationship further.

Notably, during the construction of the ML model in this study, although relatively advanced algorithms and methods were adopted, a comprehensive routine validation process was not carried out. The main reasons are as follows: This study aimed to explore new preliminary models and methods for assessing the carcinogenesis risk of CAG. There were certain limitations in data collection, and the sample size was relatively limited, which made it difficult to support a complex and comprehensive validation process. Moreover, the research time was relatively tight, and it was difficult to implement various validation methods, such as cross-validation, within a limited time.

In conclusion, this study successfully constructed a risk assessment model for the carcinogenesis of CAG. Moreover, we identified deficiencies in research on the relationships between hematological parameters and the carcinogenesis of CAG, which highlights directions for subsequent research. In the future, we will expand the sample size, improve the validation process, and explore the mechanisms of action of various factors, especially hematological parameters, in carcinogenesis risk assessment. We will further optimize the model to provide more reliable and effective tools and strategies for the early and precise assessment of the carcinogenesis of CAG and clinical intervention and promote the development of research in this field.

5 Conclusion

In conclusion, this study successfully constructed a risk assessment model for the carcinogenesis of CAG. By analyzing the medical records of patients from the Affiliated Hospital of Qingdao University, we defined strict inclusion and exclusion criteria to ensure the reliability of the research subjects. Through comprehensive evaluation methods, we aimed to assess the risk of carcinogenesis accurately. All pathological data were approved, and strict ethical standards were followed. This research not only provides new insights into the field of CAG carcinogenesis but also lays a foundation for further exploration of early warning and intervention strategies. However, given the limitations in data collection and time, future research is needed to expand the sample size, improve the validation process, and optimize the model for more accurate risk prediction.

Acknowledgements None.

Author contributions JY.Z was responsible for collecting case data and analyzing them, and also wrote the manuscript. DL and GJ.H were responsible for proofreading and revising the manuscript. Each author played a significant role in the research process, contributing to the development of the research question, the design of the study, the collection of data, the analysis of findings, and the interpretation of results. Furthermore, all authors were actively involved in the drafting, revision, and critical review of the article. The decision to submit the article to the chosen journal was made by all authors, and the final version was approved collectively. Moreover, all authors accept responsibility for the entirety of the work.

Funding There is no funding to report.

Data availability Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate This protocol was conducted in accordance with the guidelines outlined in the Declaration of Helsinki and received approval from the Ethics Committee of the Affiliated Hospital of Qingdao University (Ethics Approval Number: QYFY WZLL 26701). Given that this study is retrospective and all statistical data are anonymized, the Ethics Committee has agreed to waive the requirement for informed consent.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Yerolatsite M, Torounidou N, Gogadis A, et al. TAMs and PD-1 networking in gastric cancer: a review of the literature. Cancers. 2023;16(1):196.
2. Sun D, Lei L, Xia C, et al. Sociodemographic disparities in gastric cancer and the gastric precancerous cascade: a population-based study. Lancet Reg Health West Pac. 2022;23:100437.
3. Pimentel-Nunes P, Libânio D, Marcos-Pinto R, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy. 2019;51(4):365–88.
4. Livzan MA, Mozgovoi SI, Gaus OV, et al. Histopathological evaluation of gastric mucosal atrophy for predicting gastric cancer risk: problems and solutions. Diagnostics. 2023;13(15):2478.
5. Su X, Liu Q, Gao X, et al. Evaluation of deep learning methods for early gastric cancer detection using gastroscopic images. Technol Health Care. 2023;31(S1):313–22.
6. Matsuoka T, Yashiro M. Biomarkers of gastric cancer: current topics and future perspective. World J Gastroenterol. 2018;24(26):2818–32.
7. Alkhamis MA, Al Jarallah M, Attur S, et al. Interpretable machine learning models for predicting in-hospital and 30 days adverse events in acute coronary syndrome patients in Kuwait. Sci Rep. 2024;14(1):1243.
8. Tufail AB, Ma YK, Kaabar MKA, et al. Deep learning in cancer diagnosis and prognosis prediction: a minireview on challenges, recent trends, and future directions. Comput Math Methods Med. 2021;2021:9025470.
9. Shehab M, Abualigah L, Shambour Q, et al. Machine learning in medical applications: a review of state-of-the-art methods. Comput Biol Med. 2022;145:105458.
10. Xu W, Jiang T, Shen K, et al. GADD45B regulates the carcinogenesis process of chronic atrophic gastritis and the metabolic pathways of gastric cancer. Front Endocrinol. 2023;14:1224832.
11. Shah SC, Piazuelo MB, Kuipers EJ, et al. AGA clinical practice update on the diagnosis and management of atrophic gastritis: expert review. Gastroenterology. 2021;161(4):1325-32.e7.
12. Nagtegaal ID, Odze RD, Klimstra D, et al. The 2019 WHO classification of tumours of the digestive system. Histopathology. 2020;76(2):182–8.
13. Bai J, Huang JH, Price CPE, et al. Prognostic factors for polyp recurrence in chronic rhinosinusitis with nasal polyps. J Allergy Clin Immunol. 2022;150(2):352-61.e7.
14. Wang H, Zhang L, Liu Z, et al. Predicting medication nonadherence risk in a Chinese inflammatory rheumatic disease population: development and assessment of a new predictive nomogram. Patient Prefer Adher. 2018;12:1757–65.
15. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the hosmer-lemeshow test revisited. Crit Care Med. 2007;35(9):2052–6.
16. Huang YQ, Liang CH, He L, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. J Clin Oncol. 2016;34(18):2157–64.
17. Balachandran VP, Gonen M, Smith JJ, et al. Nomograms in oncology: more than meets the eye. Lancet Oncol. 2015;16(4):e173–80.
18. Li Y, Bao Y, Zheng H, et al. A nomogram for predicting severe myelosuppression in small cell lung cancer patients following the first-line chemotherapy. Sci Rep. 2023;13(1):17464.
19. Luan X, Zhao L, Zhang F, et al. Sex disparity, prediagnosis lifestyle factors, and long-term survival of gastric cancer: a multi-center cohort study from China. BMC Cancer. 2024;24(1):1149.
20. Qin J, Liu M, Ding Q, et al. The direct effect of estrogen on cell viability and apoptosis in human gastric cancer cells. Mol Cell Biochem. 2014;395(1–2):99–107.
21. Ji Y, Zhang Y, Yun Q, et al. Gender differences in social environmental changes associated with smoking: a cross-sectional study from Chinese internal migrants. BMJ Open. 2022;12(11): e058097.
22. Rota M, Possenti I, Valsassina V, et al. Dose-response association between cigarette smoking and gastric cancer risk: a systematic review and meta-analysis. Gastric Cancer. 2024;27(2):197–209.
23. Butt J, Varga MG, Wang T, et al. Smoking, helicobacter pylori serology, and gastric cancer risk in prospective studies from China, Japan, and Korea. Cancer Prev Res. 2019;12(10):667–74.
24. Wu H, Chen HL. The association between heavy alcohol use and gastric cancer. Am J Gastroenterol. 2021;116(12):2470–1.
25. Tas F, Ciftci R, Kilic L, et al. Clinical and prognostic significance of coagulation assays in gastric cancer. J Gastrointest Cancer. 2013;44(3):285–92.
26. Zhu L, Liu S, Wang D, et al. Relationship between coagulation and prognosis of gastric cancer: a systematic review and meta-analysis. Curr Ther Res Clin Exp. 2024;101:100741.
27. Zhu XS, Zhao Y, Ma FY, et al. Value of preoperative hematological parameters in the prognosis of gastric cancer patients undergoing a total gastrectomy. Curr Med Sci. 2022;42(2):348–56.
28. Dirican A, Ekinci N, Avci A, et al. The effects of hematological parameters and tumor-infiltrating lymphocytes on prognosis in patients with gastric cancer. Cancer Biomark. 2013;13(1):11–20.
29. Xu S, Fan Y, Tan Y, et al. Association between blood lipid levels and risk of gastric cancer: a systematic review and meta-analysis. PLoS ONE. 2023;18(7): e0288111.
30. Zhang D, Hu RH, Cui XM, et al. Lipid levels and insulin resistance markers in gastric cancer patients: diagnostic and prognostic significance. BMC Gastroenterol. 2024;24(1):373.