Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–10

DOI: 10.1177/20552076241277713
journals.sagepub.com/home/dhj

Sage

A feasibility study on utilizing machine learning technology to reduce the costs of gastric cancer screening in Taizhou, China

Si-yan Yan[1,*], Xin-yu Fu[2,*], Shen-Ping Tang[3,*], Rong-bin Qi[2], Jia-wei Liang[4], Xin-li Mao[3,5,6], Li-ping Ye[3,5,6] and Shao-wei Li[3,5,6]

Abstract

Aim: To optimize gastric cancer screening score and reduce screening costs using machine learning models.

Methods: This study included 228,634 patients from the Taizhou Gastric Cancer Screening Program. We used three machine learning models to optimize Li's gastric cancer screening score: Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), and Deep Learning (DL). The performance of the binary classification models was evaluated using the area under the curve (AUC) and area under the precision–recall curve (AUCPR).

Results: In the binary classification model used to distinguish low-risk and moderate- to high-risk patients, the AUC in the GBM, DRF, and DL full models were 0.9994, 0.9982, and 0.9974, respectively, and the AUCPR was 0.9982, 0.9949, and 0.9918, respectively. Excluding *Helicobacter pylori* IgG antibody, pepsinogen I, and pepsinogen II, the AUC in the GBM, DRF, and DL models were 0.9932, 0.9879, and 0.9900, respectively, and the AUCPR was 0.9835, 0.9716, and 0.9752, respectively. Remodel after removing variables IgG, PGI, PGII, and G-17, the AUC in GBM, DRF, and DL was 0.8524, 0.8482, 0.8477, and AUCPR was 0.6068, 0.6008, and 0.5890, respectively. When constructing a tri-classification model, we discovered that none of the three machine learning models could effectively distinguish between patients at intermediate and high risk for gastric cancer (F1 scores in the GBM model for the low, medium and high risk: 0.9750, 0.9193, 0.5334, respectively; F1 scores in the DRF model for low, medium, and high risks: 0.9888, 0.9479, 0.6694, respectively; F1 scores in the DL model for low, medium, and high risks: 0.9812, 0.9216, 0.6394, respectively).

Conclusion: We concluded that gastric cancer screening indicators could be optimized when distinguishing low-risk and moderate to high-risk populations, and detecting gastrin-17 alone can achieve a good discriminative effect, thus saving huge expenditures.

Keywords

Machine learning, gastric cancer, gastric cancer screening, gastrin-17, pepsinogen I and II, *Helicobacter pylori* antibody

Submission date: 28 January 2024; Acceptance date: 8 August 2024

[1]Taizhou Hospital of Zhejiang Province, Zhejiang University, Hangzhou, Zhejiang, China
[2]Taizhou Hospital of Zhejiang Province affiliated to Wenzhou Medical University, Linhai, Zhejiang, China
[3]Department of Gastroenterology, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, Zhejiang, China
[4]Department of Thoracic Surgery, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai Zhejiang, China
[5]Key Laboratory of Minimally Invasive Techniques & Rapid Rehabilitation of Digestive System Tumor of Zhejiang Province, Taizhou Hospital Affiliated to Wenzhou Medical University, Linhai, Zhejiang, China
[6]Institute of Digestive Disease, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, Zhejiang, China
[*]These authors contributed equally to this work.

Corresponding authors:
Xin-li Mao, Key Laboratory of Minimally Invasive Techniques & Rapid Rehabilitation of Digestive System Tumor of Zhejiang Province, Taizhou Hospital Affiliated to Wenzhou Medical University, Linhai, Zhejiang, 317000, China.
Email: maoxl@enzemed.com
Li-ping Ye, Taizhou Hospital of Zhejiang Province affiliated to Wenzhou Medical University, Linhai, Zhejiang, China.
Email: yelp@enzemed.com
Shao-wei Li, Department of Gastroenterology, Taizhou Hospital of Zhejiang Province affiliated to Wenzhou Medical University, Linhai, Zhejiang, China.
Email: li_shaowei81@hotmail.com

Introduction

Gastric cancer is among the most prevalent forms of cancer worldwide and represents a high-incidence area.[1] According to data from the World Health Organization, in 2020, there were 479,000 newly diagnosed cases of gastric cancer in our country, with 374,000 deaths, accounting for 44.0% and 48.6% of new global cases and deaths, respectively.[2] The incidence and mortality rates of gastric cancer remain the highest in Asia, posing a severe threat to the lives and health of our people and imposing significant economic burdens on society and families.[3] Given that early-stage gastric cancer lacks specific clinical symptoms, the majority of patients are already in the advanced or late stages by the time of the diagnosis,[4] resulting in a poor prognosis. Therefore, the early diagnosis and treatment are vital to reduce the mortality rate of gastric cancer. However, the early detection rate of gastric cancer in our country is low, at <10%,[5] which is far below South Korea's 50% and Japan's 70%.[6] In addition, there is no internationally recognized screening program for gastric cancer, and early detection relies mainly on opportunistic screening.[7] Thus, it is worth exploring how to economically and effectively implement a gastric cancer screening strategy that suits national conditions.

According to a study conducted in the United States, gastric cancer screening is cost-effective in countries with high incidence rates and low screening costs, and it may still be a feasible option for high-risk populations in countries with low incidence rates.[8] In China, over 300 million individuals are estimated to be at risk of gastric cancer.[9] Although gastroscopy and gastric biopsies are currently considered the gold standards for screening and diagnosing gastric cancer, they are expensive and invasive procedures, leading to poor patient compliance and tolerance.[10] Therefore, conducting gastroscopic screening for such a large population is inefficient and impractical, considering that only 1%–3% of individuals are expected to have gastric cancer.[11] Therefore, further risk stratification of this at-risk population, as a preliminary screening before gastroscopy, is necessary to identify high-risk individuals among them.

Li's gastric cancer screening score is currently employed for prospective gastric cancer screening in asymptomatic individuals, aiming to differentiate low-, moderate-, and high-risk populations, with gastroscopy recommended for those at moderate to high risk.[11,12] This approach has been validated in multicenter studies focusing on the Chinese population.[11] However, the initial screening based on Li's score requires consideration of four serological markers: gastrin-17 (G-17), pepsinogen I and II (PGI/II), and *Helicobacter pylori* antibody (Hp IgG), rendering it cost-intensive.

Given the above, the present study aimed to optimize Li's score by reducing the initial screening markers and lowering the cost, thus enhancing the cost-effectiveness of gastric cancer screening. In recent years, machine learning has demonstrated significant potential in data processing and analyses. Predictive models built on machine learning exhibit excellent predictive performance and stability, garnering increasing attention. By leveraging machine learning and a comparative analysis of multiple models, this study optimized the Li's score. Compared to the original Li's score, these models reduce the need for blood markers, substantially decreasing the screening cost while maintaining a discrimination effect close to 100% in both internal and external validation, ensuring excellent screening efficacy.

Methods

Trial population

From January to December 2019, the participants in the gastric cancer screening program for the minimum guaranteed population in the Taizhou area and the participants in the gastric cancer screening program for the general population in Linhai City were recruited from January 2021 to July 2023. Both screening programs were approved by the Ethics Committee of Taizhou Hospital Affiliated with Wenzhou Medical University, with ethics approval numbers (K20190221 and K20210613).

The inclusion criteria were all participants in the screening project, while exclusion criteria were as follows: refusal to complete the questionnaire survey, refusal to undergo serological testing, and missing data. All participants provided written informed consent after fully understanding the benefits and risks of GC screening.

The questionnaire survey included information such as the age, sex, height, weight, medical history, alcohol consumption history, and smoking history. Serological testing included measurements of G-17, PGI/II, and Hp IgG. Serological testing was conducted by a designated company appointed by the Taizhou Municipal Government to ensure the stability of the test results. The initial screening process utilized Li's gastric cancer screening score (Supplemental Table S1).

Data analyses

Model selection. The research involved the selection of three models: the Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), and Deep Learning (DL) models. The GBM is a machine-learning method that builds a powerful model by iteratively training multiple weak learners. It possesses the advantages of high accuracy and adaptability to different data types but has a longer training time and is susceptible to the influence of noise and outliers.[13] The DRF is a distributed implementation of the Random Forest, which is suitable for handling large-scale datasets. It can utilize distributed computing resources to obtain more accurate results through parallel computations.[14] In contrast, the DL model is a machine learning approach based on neural networks and

characterized by multilayer network structures.[15] It has a powerful learning ability and is suitable for large-scale data. In addition we appended 10 machine learning models for further explanation of the feasibility of the study, and the results are appended in Supplemental Table S2 and Figures S1–S40.

Evaluating the models. For model validation, we employed both internal and external validation. For internal validation, we used a five-fold cross-validation approach. In the evaluation of the model, we adopted the following metrics: accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUCPR). The interpretation of the model was analyzed using SHAP values. Model construction and data analyses were performed using the R language (version 4.2.3).

Results

Patients

A total of 240,059 patients participated in the project, with 11,425 patients excluded because of missing or incomplete blood test information, resulting in a final inclusion of 228,634 patients. The training set consisted of 195,640 individuals from the general population with a median age of 60 years old, including 78,539 males and 117,101 females. Among them, 34,135 had a history of smoking, 26,672 had a history of alcohol consumption, 43,070 had a history of hypertension, 13,454 had a history of diabetes, and 5651 had a history of hyperlipidemia. According to the risk assessment form, 147,522 were classified as low risk, 45,222 as medium risk, and 2896 as high risk. Those classified as medium-to high risk were designated as the endoscopy group, while the low-risk group was designated as the follow-up group. Ultimately, 48,118 individuals were included in the endoscopy group, requiring further in endoscopic examinations, while 147,522 patients were included in the follow-up group. In addition, 32,994 individuals from the low-income group were selected for validation, with a median age of 58 years old (20,107 males and 12,887 females). Among them, 10,171 had a history of smoking, 7710 had a history of alcohol consumption, 8276 had a history of hypertension, 2244 had a history of diabetes, and 2210 had a history of hyperlipidemia. Ultimately, there were 22,386 individuals in the follow-up group and 10,608 in the endoscopy group. The baseline distributions of the training and validation sets are listed in Table 1.

Constructing a binary classification model

All variables were included in the scoring table, and full models were built using GBM, DRF, and DL. In the GBM full model, the AUC was 0.9994, AUCPR was 0.9982. External validation results: recall is 0.9679, precision is 0.9351, F1 score is 0.9512, accuracy is 0.9681, AUC is 0.9959 and AUCPR is 0.9923. In the DRF full model, the AUC and AUCPR were both 0.9949. External validation results: recall is 0.9228, precision is 0.9494, F1 score is 0.9359, accuracy is 0.9594, AUC is 0.9896 and AUCPR is 0.9864. In the DL full model, the AUC was 0.9974, AUCPR was 0.9918. External validation results: recall is 0.9315, precision is 0.9601, F1 score is 0.9456, accuracy is 0.9201, AUC is 0.9912 and AUCPR is 0.9804 (Supplemental Figures S41–S43). For these three machine learning models, we used SHAP values for interpretation and the results are shown in Figure 1.

Variable selection. Based on the ranking of feature importance across the three models, IgG had the lowest weight in the GBM model, followed by PGI and PGII. In the DRF model, IgG had the least weight, followed by PGI and PGII, whereas in the DL model, PGII had the least weight, followed by PGI and IgG. The importance of the variables' features was reevaluated within the models by rearranging the variables using the permutation method. In the GBM model, the least impactful variable is PGII, followed by IgG and PGI; in the DRF model, PGII has the lowest importance weight, followed by PGI and IgG; in the DL model, IgG had the lowest weight, followed by PGII and PGI. It was observed that PGI, PGII, and IgG had relatively low weights in all three machine models. Furthermore, the models were reconstructed by removing the IgG and PGR variables (PGI/PGII) (Supplemental Figures S44–S49). Based on this, we used Partial Dependence Plot to show the average effect of the first three important feature variables in the model on the target variables, and the results are shown in Supplemental Figures S50–S58.

Removal of variables to construct models. After removing the variable IgG, GBM, DRF, and DL models were rebuilt. In the GBM model, the AUC was 0.9979, and the AUCPR was 0.9939. External validation results demonstrated a recall of 0.9818, a precision of 0.9546, an F1 score of 0.9680, an accuracy of 0.9791, an AUC of 0.9947, and an AUCPR of 0.9893. In the DRF model, the AUC was 0.9964, accompanied by an AUCPR of 0.9894. External validation results demonstrated a recall of 0.9734, a precision of 0.9525, an F1 score of 0.9628, an accuracy of 0.9758, an AUC of 0.9904, and an AUCPR of 0.9857. The DL model exhibited an AUC of 0.9931 and AUCPR of 0.9802. External validation results demonstrated a recall of 0.9570, a precision of 0.9447, an F1 score of 0.9507, an accuracy of 0.9681, an AUC of 0.9869, and an AUCPR of 0.9770 (see Supplemental Figures S59–S61).

After removing the PGI and PGII variables, in the GBM model, the AUC was 0.9961, and the AUCPR was 0.9912. External validation results demonstrated a recall of 0.9634, a precision of 0.9645, an F1 score of 0.9640, an accuracy of 0.9768, an AUC of 0.9948, and an AUCPR of 0.9913. In the DRF model, the AUC was 0.9920, and the AUCPR

Table 1. Baseline comparison table of MLGC with the general population.

	MLGC $N = 32,994$	General population $N = 195,640$	
			p-value
Age	58 (52,64)	60 (54,66)	<.001
Gender			<.001
Male	20,107 (62.26%)	78,539 (40.14%)	
Female	12,887 (39.06%)	117,101 (59.86%)	
Smoking history			<.001
No	22,823 (69.17%)	161,505 (82.55%)	
Yes	10,171 (30.83%)	34,135 (17.45%)	
Drinking history			<.001
No	25,284 (76.63%)	168,968 (86.37%)	
Yes	7710 (23.37%)	26,672 (13.63%)	
Past medical history			<.001
Hypertensive disease	8276 (25.08%)	43,070 (22.01%)	
Diabetes	2244 (6.80%)	13,454 (6.88%)	
Hyperlipidemia	2210 (6.70%)	5651 (2.89%)	
Modeling subgroups			<.001
Follow-up group	22,386 (67.85%)	147,522 (75.40%)	
Endoscopy group	10,608 (32.15%)	48,118 (24.60%)	
Gastric cancer risk class			<.001
Low-risk	22,386 (67.85%)	147,522 (75.40%)	
Medium-risk	10,045 (30.44%)	45,222 (23.11%)	
High-risk	563 (1.71%)	2896 (1.48%)	

MLGC: minimum living guarantee crowds; BMI: Body Mass Index.

was 0.9840. External validation results demonstrated a recall of 0.9268, a precision of 0.9872, an F1 score of 0.9560, an accuracy of 0.9726, an AUC of 0.9936, and an AUCPR of 0.9890. In the DL model, the AUC was 0.9921, and the AUCPR was 0.9828.External validation results demonstrated a recall of 0.9061, a precision of 0.9508, an F1 score of 0.9279, an accuracy of 0.9654, an AUC of 0.9888, and an AUCPR of 0.9777(see Supplemental Figures S62–S64).

After removing the IgG, PGI, and PGII variables, the GBM model achieved an AUC of 0.9932 and an AUCPR of 0.9835. External validation results demonstrated a recall of 0.9804, a precision of 0.9528, an F1 score of 0.9664, an accuracy of 0.9781, an AUC of 0.9934, and an AUCPR of 0.9876. In the DRF model, AUC and AUCPR were 0.9879 and 0.9716, respectively. External validation results demonstrated a recall of 0.9638, a precision of 0.9424, an F1 score of 0.9529, an accuracy of

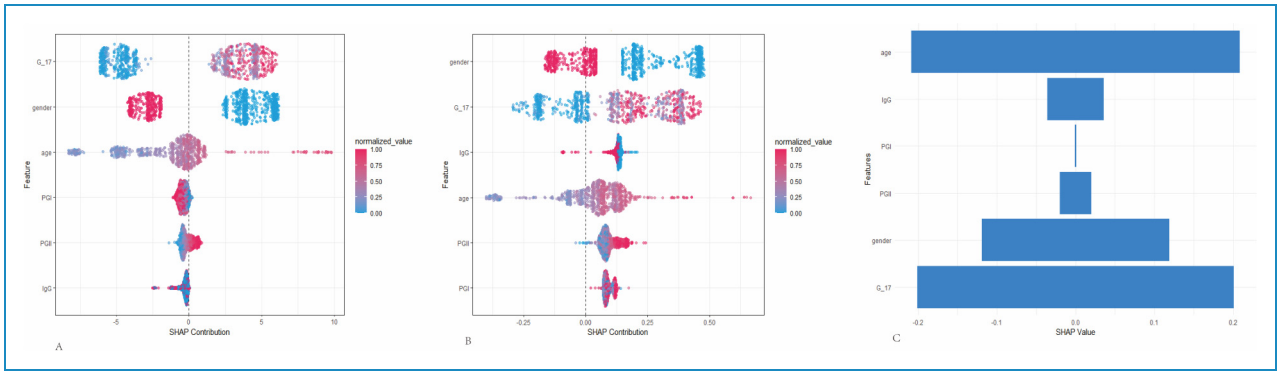

Figure 1. Plot of SHAP values for the three models; (a) distribution of SHAP values for the GBM model; (b) distribution of SHAP values for the DRF model; (c) distribution of SHAP values for the DL model (the DL model could not be mapped by the tree-based algorithm, so a separate calculation was performed).
GBM: Gradient Boosting Machine; DRF: Distributed Random Forest; DL: Deep Learning.

0.9694, an AUC of 0.9902, and an AUCPR of 0.9826.For the DL model, the AUC was 0.9900 with an AUCPR of 0.9752. External validation results demonstrated a recall of 0.9638, a precision of 0.9424, an F1 score of 0.9530, an accuracy of 0.9694, an AUC of 0.9886, and an AUCPR of 0.9722 (see Supplemental Figures S65–S67).

Constructing a tri-classification classification model

In the creation of the tri-classification model within the comprehensive GBM model, the accuracy was calculated based on the confusion matrix and yielded a value of 0.9859. The recall rates for the low-, medium-, and high-risk categories were 0.9958, 0.9665, and 0.7832, respectively. The precision rates for the low-, medium-, and high-risk categories were 0.9906, 0.9723, and 0.9497, respectively. The F1 scores for the low-, medium-, and high-risk categories were 0.9932, 0.9694, and 0.8488, respectively. For the external validation, the accuracy was determined to be 0.9524. The recall rates for the low-, medium-, and high-risk categories were 0.9979, 0.8908, and 0.3364, respectively. The precision rates for the low-, medium-, and high-risk categories were 0.9531, 0.9496, and 0.9955, respectively. The F1 scores for the low-, medium-, and high-risk categories were 0.9750, 0.9193, and 0.5334, respectively.

In the DRF model, the accuracy was 0.9760. The recall rates for the low-, medium-, and high-risk categories were 0.9918, 0.9445, and 0.6598, respectively. The precision rates for the low-, medium-, and high-risk categories were 0.9859, 0.9514, and 0.8156, respectively. The F1 values for the low-, medium-, and high-risk categories were 0.9888, 0.9479, and 0.6694, respectively. For the validation set, the accuracy was 0.9521. The recall rates for the low-, medium-, and high-risk categories were 0.9979, 0.8909, and 0.3227, respectively. The precision rates for the low-, medium-, and high-risk categories were 0.9532, 0.9484, and 0.9953, respectively. The F1 values for the low-, medium-, and high-risk categories were 0.9750, 0.9188, and 0.5175, respectively.

In the DL model, the accuracy was 0.9630. The recall rates for the low-, medium-, and high-risk categories were 0.9740, 0.9396, and 0.7826, respectively. The precision rates for the low-, medium-, and high-risk categories were 0.9885, 0.9042, and 0.6961, respectively. The F1 values for the low-, medium-, and high-risk categories were 0.9812, 0.9216, and 0.6394, respectively. In the external validation, the accuracy was 0.9468. The recall rates for the low-, medium-, and high-risk categories were 0.9862, 0.8955, and 0.3864, respectively. The precision rates for the low-, medium-, and high-risk categories were 0.9598, 0.9271, and 0.6623, respectively. The F1 values for the low-, medium-, and high-risk categories were 0.9728, 0.9110, and 0.3945, respectively. The parameter distributions of the three models are shown in Figure 2.

Discussion

As a country with a high incidence of gastric cancer, it is necessary for China to conduct gastric cancer screening. However, a standardized system for gastric cancer screening has not yet been established in China, and screening still relies on opportunistic approaches, leading to a high incidence and mortality rate of gastric cancer in China, ranking top in Asia. However, large-scale gastric cancer screening in China is expensive, so it is necessary to stratify the population for initial screening.

Li's Gastric Cancer Screening Form, developed and validated for the Chinese population, was used to identify high-risk groups, with the aim of reducing screening costs. Can machine-learning algorithms be used to further optimize the screening form and reduce screening costs? Therefore, this study, relying on the gastric cancer screening project of the Taizhou Municipal Government in

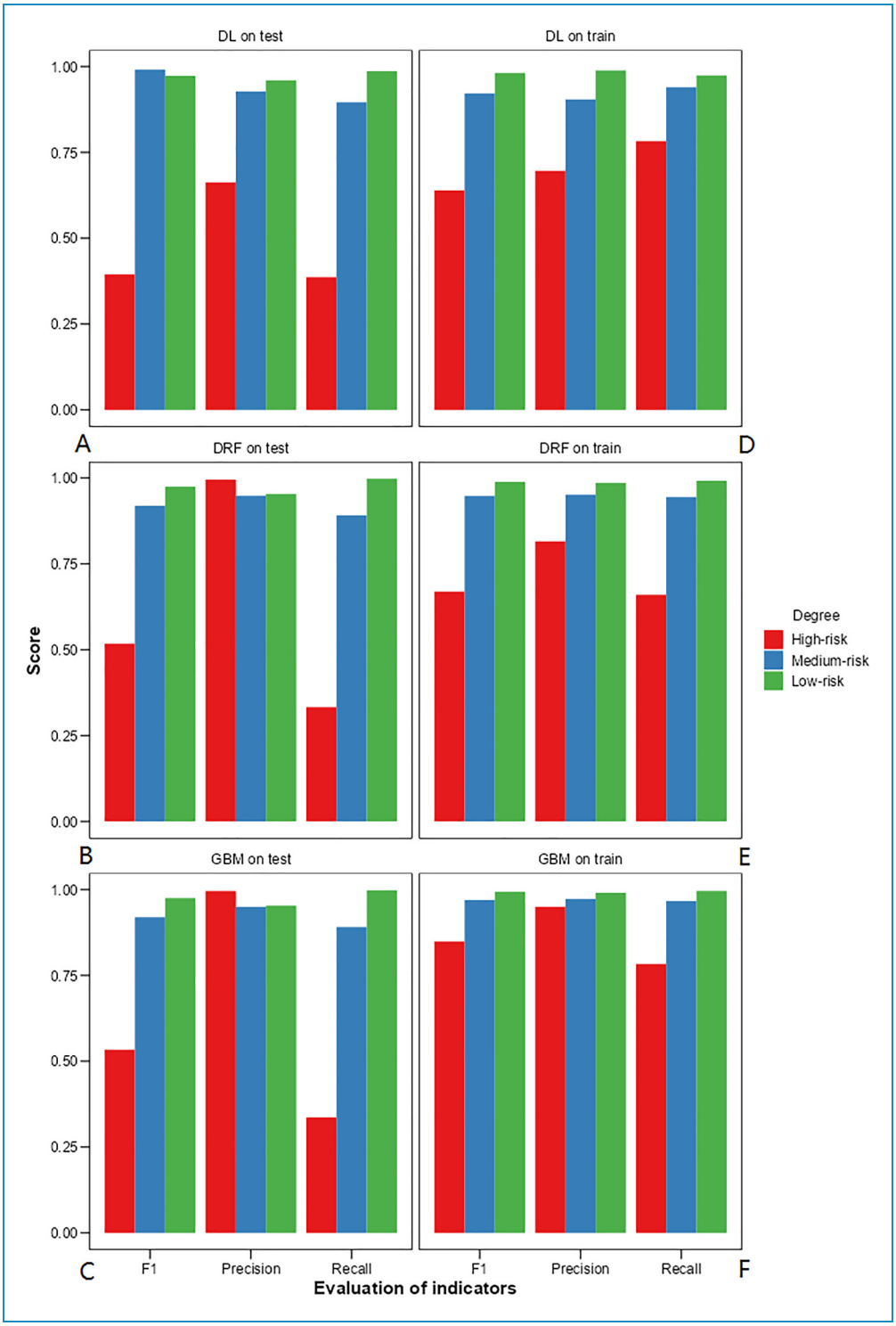

Figure 2. The evaluation of the tri-classification model to distinguish patients at low, medium, and high risk. (a), (b), and (c) show the F1 value, precision rate, and recall rate of the DL, DRF, and GBM models on the test set. (d), (e), and (f) show the F1 value, precision rate, and recall rate of the DL, DRF, and GBM models on the training set.
GBM: Gradient Boosting Machine; DRF: Distributed Random Forest; DL: Deep Learning.

China and based on a large sample of data, employed various machine learning models to optimize the gastric cancer screening form. When simply distinguishing between low-risk and moderate-to-high-risk individuals, the screening indicator gastrin alone can achieve a 99% discrimination rate.

This study employed three machine-learning algorithms (GBM, DRF, and DL) to construct a discriminative model encompassing all variables. The models demonstrated outstanding discriminatory performance in both internal and external validation. To explore the significance of the variables within the models, a visual representation of the variables was conducted, and internal variable reordering was performed to further assess their weights. Through variable analyses, it was observed that IgG, PGI, and PGII exhibited relatively minor weights in all three models. As a result, we hypothesized that these indicators could potentially be optimized. Thus, the study proceeded to remove IgG and recreate the model. This adjustment resulted in a slight decrease in discriminative performance, albeit maintaining a good predictive effect. Subsequently, PGI and PGII were eliminated to construct the discriminative model. Although this resulted in a mild reduction in discriminative performance compared to the full model, it was comparable to the model with only IgG removed. Furthermore, an attempt was made to exclude IgG, PGI, and PGII and solely include the age, gender, and G-17 in the machine learning model. Across the three models, this led to varying degrees of decrease in discriminative performance. The GBM model exhibited the least impact, with an AUC of 0.9925 and an AUC of 0.9934 and an AUCPR of 0.9876 in the external validation. The overall discriminative effect was close to 99%. Notably, removing G-17 significantly decreased the overall discriminative effect. Based on iterative modeling and validation, it was concluded that the GBM model outperforms the other models, displaying excellent discriminative performance with minimal impact on variable removal. Consequently, we believe that for gastric cancer screening in the general population, G-17 alone can effectively differentiate between low-risk and medium-to-high-risk individuals. This reduction from four initial serological markers to one substantially lowers the costs in large-scale gastric cancer screening programs. Although IgG, PGI, and PGII help improve discrimination, they are not the primary determinants.

For binary classification, it has been observed that the indicator G-17 alone is sufficient to effectively differentiate between individuals who need to undergo gastroscopy and those who do not. G-17 carried a significant weight in the model, whereas the weights of IgG, PGI, and PGII had a minor impact on the model. The possible reasons for this phenomenon are described below.

The presence of *H. pylori* IgG antibodies can only indicate past infection in the body. IgG antibodies can still be detected several months after treatment, which can potentially interfere with gastric cancer screening.[16] In addition, many studies have demonstrated that IgG antibody testing has a high negative predictive value for detecting active infections. The efficacy of this test in detecting active infections depends on factors such as the patient's age, dietary patterns, clinical conditions of the infection, and selection of antigens used for antibody preparation in the enzyme-linked immunosorbent assay kit.[17–19] A retrospective study conducted in Japan indicated that IgG antibodies can, to some extent, indicate the presence of gastric mucosal damage, although their accuracy is influenced by age.[20] In patients over 65 years old, the reaction of IgG antibodies to *H. pylori* is diminished compared with younger patients, making it impossible to indicate the presence of infection. Furthermore, there are reports suggesting that the accurate reflection of *H. pylori* infection status and treatment effects may require the dynamic measurement of antibody titers.[21] In addition, in the scoring system, the value assigned to *H. pylori* is only one point, indicating its limited role in differentiation.[22]

G-17, released from G cells, is an important gastrointestinal peptide hormone that reflects secretion in the antrum.[23] G-17 plays a role in the occurrence and development of GC by influencing inflammatory processes and gut microbiota.[24] The antrum of the stomach is widely recognized as a high-risk area for gastric cancer. The levels of G-17 are related to the degree of gastric mucosal atrophy and the type, location, and extent of gastric cancer. Relevant studies have shown a close correlation between G-17 and gastric cancer, as it can reflect malignant transformations and abnormal proliferation.[25,26] Elevated levels of gastrin promote the development of gastric cancer by inhibiting apoptosis of cancer cells.[25] When the gastric region is damaged for various reasons, the gastric mucosal function undergoes varying degrees of change. As the gastric mucosa progresses from superficial lesions to atrophy, gastric ulcers, and carcinogenesis, the serum level of G-17 progressively increases. Previous studies have shown that G-17 levels increase successively in nonatrophic gastritis, early gastric cancer, and advanced gastric cancer,[27,28] indicating a possible correlation between the G-17 expression and disease stages.

PGI and PGII are also considered important markers of chronic atrophic gastritis and are primarily secreted by the main cells of the gastric body.[29] Furthermore, research has indicated a positive correlation between G-17 and PGI/PGII levels. When G-17 levels increase, PGI and PGII levels may also increase accordingly.[30] This suggests a correlation among PGI, PGII, and G-17, where G-17 reflects PGI and PGII levels to some extent. This may be due to the fact that gastric inflammation is not limited to a specific region. For example, when antral gastritis is present, G-17 levels significantly increase; however, as a result, it also stimulates the gastric body, leading to increased secretion of PGI and PGII. Lesions in the gastric region, whether in the antrum or body, affect the levels of G-17, PGI, and PGII. In our model constructed using the dataset, the weights assigned to PGI and PGII were significantly lower than those assigned to G-17. This may be due to the fact that our dataset is from a single center in the Taizhou region, where G-17 levels are

generally higher than PGI and PGII levels, resulting in a higher weight in the model. Furthermore, studies have indicated a close correlation between G-17 and *H. pylori*, where G-17 levels are significantly higher in patients with *H. pylori* infection than in those without it.[31,32] This also suggests that G-17 can, to some extent, reflect *H. pylori* infection. Based on the above discussion, we concluded that removing IgG, PGI, and PGII from the model had a small effect on the classification effectiveness of the model. However, considering the cost savings achieved through this screening, the cost remained within an acceptable range.

Up to now, there have been numerous studies on using serological markers such as G-17, PGI, PGII, and Hp-IgG for gastric cancer screening. Previous studies have shown that in screening GC, the diagnostic value of a single serological marker was not good, and the combination of different markers can improve diagnostic efficiency.[33] In Japan, a country known to have a high prevalence of GC, the ABC method using Hp-IgG and serum pepsinogen (PG) has been used in large-scale GC screening with satisfactory results.[34] However, studies have indicated that the ABC method has a relatively low overall detection rate of GC in China, and its sensitivity and specificity are not as good as reported abroad. Some scholars proposed a new ABC method in China using G-17 combined with PG as serological screening indicators. Studies have proven the effectiveness of the new ABC method.[33,35] Notably, in areas with a low incidence of GC and Hp, Ghoshal et al. found that PGR and G-17 tests were not good predictors of gastric precancerous lesions.[36] These differences might be attributed to racial, environmental, dietary, and socioeconomic factors. These findings suggest that detection methods suitable for different regions may not be universally applicable. Policymakers should fully consider the characteristics of countries and regions and develop gastric cancer screening strategies suitable for local conditions to improve the efficiency of GC screening programs. Our study utilized multiple machine learning models to uniquely highlight the cost-effectiveness of using G-17 alone for initial screening in the Taizhou area, reducing costs significantly while maintaining high discriminative performance. Previous studies have not extensively focused on this aspect of cost optimization through serological marker reduction. Although the region we included may not fully represent China, our study could still provide significant insights for policymakers.

When constructing a tri-classification model, we discovered that none of the three machine learning models was able to effectively distinguish individuals at intermediate and high risk for gastric cancer. We attribute this to the disproportionately small proportion of high-risk individuals within such a large training sample, which significantly influences the accuracy of the model. Furthermore, for large-scale screening purposes, the clinical significance of distinguishing between high- and intermediate-risk individuals is minimal. While the recommendation for gastroscopy differs for individuals at intermediate and high risk, both groups would still be advised to undergo a thorough endoscopic examination during screening. Therefore, we believe that the binary classification model has greater clinical relevance and applicability. Our study's discriminatory model has been extensively trained and validated using a large sample size, and holds clinical value in the Taizhou region. After removing the screening indicators IgG, PGI, and PGII, the GBM model's decrease in its discriminatory effect was within 1%.

Health economics evaluation: From the perspective of screening initiatives, more than 30 million RMB including the cost of propaganda, personnel service, materials, endoscopic examination, and pathological biopsy has been allocated to the free gastric cancer screening program for the minimum guaranteed population in the Taizhou area and the general population in Linhai City. Based on the company's testing fees, the detection costs of Hp-IgG and PGI/PGII were 30 RMB and 160 RMB per person respectively. However, the G-17 test only costs 25 RMB. The total cost of initial serum testing is 215 RMB per person. The cost of an endoscopic examination is determined uniformly within Taizhou City, being set at 800 RMB per person. Therefore, based on the currently completed screening population, excluding IgG and PGI/PGII testing can save 45.611210 million RMB. These savings could cover an additional 212,145 initial screenings by the original Li's score. Based on the results of our previous 30-month gastric cancer screening project conducted among the general population in Linhai City, 43 cases of gastric cancer were detected out of 20,456 individuals. Currently, the detection rate of the GC screening program in Linhai City is 0.21%. By extending the coverage to a further population of 212,145 individuals, approximately 445 more patients with GC could be detected. Moreover, the surplus costs obtained by implementing the optimized model could provide an additional 57,014 individuals with endoscopic examinations. According to relevant literature, GC screening is cost-effective in countries with a high incidence of GC. By improving screening techniques, it is possible to further reduce screening costs and achieve higher quality-adjusted life-years with a lower incremental cost-effectiveness ratio.

According to the survey, there are over 300 million high-risk individuals for gastric cancer in China. Due to the substantial costs involved, it is challenging to perform endoscopic examinations on all patients. The optimized gastric cancer screening scoring system in this study can achieve 99% accuracy in determining the necessity of endoscopy by solely examining G-17 levels. This approach can save nearly 57 billion RMB in screening costs, which is quite significant. The funds saved from the screening can help treat patients with advanced-stage GC or expand the screening scope to detect more early-stage cancer cases,

which can help reduce subsequent treatment expenses. Therefore, when conducting gastric cancer screening in the general population, the optimized screening scale can greatly reduce the cost of initial screening, thereby covering more people and detecting more early gastric cancers. These results will contribute to the development and implementation of nationwide gastric cancer screening programs.

Several limitations associated with the present study warrant mention. First, our study included individuals from Taizhou City, Zhejiang province. The region was chosen due to its high incidence rates of gastric cancer. The incidence of gastric cancer in Taizhou City in 2019 was 24.98/100,000. Although our study provided significant insights, the region included may not fully represent the entire Chinese population. This is because the incidence of gastric cancer varies across different regions in China due to genetic, environmental, dietary, and socioeconomic factors. The incidence is significantly higher in the northwest and eastern coastal regions. Since our study was conducted in a high-incidence area, the findings may be particularly relevant to similar high-risk regions but may not fully extrapolate to areas with lower gastric cancer incidence, thus affecting the generalizability of our results. Therefore, the conclusions of this study need to be validated through multicenter or prospective studies. Additionally, the external validation for this study also came from the Taizhou population data. Although the model demonstrated good stability and predictive performance, further validation using external data from different regions is necessary to optimize its applicability. Secondly, levels of pepsinogen (PG) and gastrin-17 (G-17) can be influenced by proton pump inhibitors (PPIs). PPIs are among the most widely distributed medications in China, with the market size for PPIs reaching nearly 30 billion RMB in 2021. Our questionnaire did not account for PPI usage, nor did we exclude patients who used antacids in the past two weeks, which may introduce some bias into the results. Thirdly, the screening project involved multiple hospitals in Taizhou. Although the screening indicators were tested by designated institutions, the sample collection and handling methods inevitably varied among different hospitals, which could also lead to some bias. Fourthly, our study included individuals aged 45–70 years. According to the 2018 data from the National Central Cancer Registry, the incidence of gastric cancer remains low before the age of 40 but rises sharply thereafter.[37] Additionally, considering China's aging population, the Chinese Guidelines for Gastric Cancer Screening and Early Diagnosis and Treatment (2022, Beijing) recommend ending screening at the age of 75. Our study may not fully cover the high-risk population recommended by the guidelines, and the results may not comprehensively reflect the actual situation of the entire recommended screening population. Future research should consider including individuals aged 40–45 and 70–75. However, expanding the screening age range, while covering more high-risk individuals, also increases the overall screening cost. Further research is needed to determine which screening range maximizes public health benefits. Despite these limitations, the present study still offers valuable insights for policymakers, particularly in regions with similar gastric cancer risk.

Conclusions

We have optimized the gastric cancer screening model using machine learning algorithms, and it has performed well in identifying low-risk and medium-high-risk individuals for gastric cancer before endoscopy examinations in the Taizhou region. By simplifying serological markers to optimize cancer screening, we have reduced the screening costs, making it highly suitable for broader implementation. Therefore, we believe that machine learning can serve as an accurate and cost-effective preliminary prescreening tool for large-scale cancer screening in Taizhou, thereby improving the detection of GC. These findings could have significant implications for policymakers and help allocate more resources to other critical aspects of cancer prevention and control.

Author contributions: XL M, SW L, XY F, and LP Y participated in Gastric Cancer Screening Program. XY F, SY Y, SP T, and RB Q participated in machine learning algorithm analysis. XY F, SW L, SY Y, RB Q, and SP T undertook validation, writing, review, and editing. All authors contributed to the article and approved the submitted version.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Open Project Program of Key Laboratory of Minimally Invasive Techniques & Rapid Rehabilitation of Digestive System Tumor of Zhejiang Province, Medical Science and Technology Project of Zhejiang Province, Major Research Program of Taizhou Enze Medical Center Grant, Program of Taizhou Science and Technology Grant (Grant Nos. 21SZDSYS01, 2024KY1788, 19EZZDA2, and 23ywa33).

ORCID iD: Shao-wei Li https://orcid.org/0000-0002-3276-1037

Supplemental material: Supplemental material for this article is available online.

References

1. Etemadi A, Safiri S and Sepanlou SG, The global, regional, and national burden of stomach cancer in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet Gastroenterol Hepatol*, 2020, 5: 42–54.

2. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; 71: 209–249.
3. Shen L, Shan YS, Hu HM, et al. Management of gastric cancer in Asia: resource-stratified guidelines. *Lancet Oncol* 2013; 14: e535–e547.
4. Zeng H, Ran X, An L, et al. Disparities in stage at diagnosis for five common cancers in China: q multicentre, hospital-based, observational study. *Lancet Public Health* 2021; 6: e877–ee87.
5. Zeng H, Chen W, Zheng R, et al. Changing cancer survival in China during 2003-15: a pooled analysis of 17 population-based cancer registries. *Lancet Glob Health* 2018; 6: e555–ee67.
6. Arnold M, Abnet CC, Neale RE, et al. Global burden of 5 major types of gastrointestinal cancer. *Gastroenterology* 2020; 159: 335–49.e15.
7. Nie Y, Wu K, Yu J, et al. A global burden of gastric cancer: the major impact of China. *Expert Rev Gastroenterol Hepatol* 2017; 11: 651–661.
8. Ascherman B, Oh A and Hur C. International cost-effectiveness analysis evaluating endoscopic screening for gastric cancer for populations with low and high risk. *Gastric Cancer* 2021; 24: 878–887.
9. He J, Chen WQ, Li ZS, et al. China guideline for the screening, early detection and early treatment of gastric cancer (2022, Beijing). *Zhonghua Zhong Liu Za Zhi* 2022; 44: 634–666.
10. Xu H and Li W. Early detection of gastric cancer in China: progress and opportunities. *Cancer Biol Med* 2022; 19: 1622–1628.
11. Cai Q, Zhu C, Yuan Y, et al. Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: a nationwide multicentre study. *Gut* 2019; 68: 1576–1587.
12. Liu K, Qin M and Huang J. The prescreening tool for gastric cancer in China. *Gut* 2020; 69: 1.
13. Zhang Z, Zhao Y, Canes A, et al. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med* 2019; 7: 52.
14. Mancini T, Calvo-Pardo H and Olmo J. Extremely randomized neural networks for constructing prediction intervals. *Neural Netw* 2021; 144: 113–128.
15. Kriegeskorte N and Golan T. Neural network models and deep learning. *Curr Biol* 2019; 29: R231–R2r6.
16. Suerbaum S and Michetti P. *Helicobacter pylori* infection. *N Engl J Med* 2002; 347: 1175–1186.
17. Xia Y, Meng G, Zhang Q, et al. Dietary patterns are associated with *Helicobacter pylori* infection in Chinese adults: a cross-sectional study. *Sci Rep* 2016; 6: 32334.
18. Yordanov D, Boyanova L, Markovska R, et al. Influence of dietary factors on *Helicobacter pylori* and CagA seroprevalence in Bulgaria. *Gastroenterol Res Pract* 2017; 2017: 9212143.
19. Mnich E, Kowalewicz-Kulbat M, Sicińska P, et al. Impact of *Helicobacter pylori* on the healing process of the gastric barrier. *World J Gastroenterol* 2016; 22: 7536–7558.
20. Toyoshima O, Nishizawa T, Sakitani K, et al. Serum anti-*Helicobacter pylori* antibody titer and its association with gastric nodularity, atrophy, and age: a cross-sectional study. *World J Gastroenterol* 2018; 24: 4061–4068.
21. Sabbagh P, Mohammadnia-Afrouzi M, Javanian M, et al. Diagnostic methods for *Helicobacter pylori* infection: ideals, options, and limitations. *Eur J Clin Microbiol Infect Dis* 2019; 38: 55–66.
22. Diseases NCCFD, Endoscopy CSOD and Association HMAOCM. China consensus on the protocol of early gastric cancer screening (Shanghai, 2017). *Chin J Gastroenterol* 2018; 23: 92–97.
23. Schubert ML. Functional anatomy and physiology of gastric secretion. *Curr Opin Gastroenterol* 2015; 31: 479–485.
24. Zhou G and Yang J. Correlations of gastrointestinal hormones with inflammation and intestinal flora in patients with gastric cancer. *J Buon* 2019; 24: 1595–1600.
25. Watson SA, Grabowska AM, El-Zaatari M, et al. Gastrin - active participant or bystander in gastric carcinogenesis? *Nat Rev Cancer* 2006; 6: 936–946.
26. Song DH, Rana B, Wolfe JR, et al. Gastrin-induced gastric adenocarcinoma growth is mediated through cyclin D1. *Am J Physiol Gastrointest Liver Physiol* 2003; 285: G217–G222.
27. Wang Y, Zhu Z, Liu Z, et al. Diagnostic value of serum pepsinogen I, pepsinogen II, and gastrin-17 levels for population-based screening for early-stage gastric cancer. *J Int Med Res* 2020; 48: 300060520914826.
28. Zheng C, Jiang Q, Wang K, et al. Nanozyme enhanced magnetic immunoassay for dual-mode detection of gastrin-17. *Analyst* 2022; 147: 1678–1687.
29. Gantuya B, Oyuntsetseg K, Bolor D, et al. Evaluation of serum markers for gastric cancer and its precursor diseases among high incidence and mortality rate of gastric cancer area. *Gastric Cancer* 2019; 22: 104–112.
30. Zeng J, Shen Y, Xu S, et al. Analysis of gastrin-17 and its related influencing factors in physical examination results. *Immun Inflamm Dis* 2023; 11: e993.
31. Leung WK, Wu MS, Kakugawa Y, et al. Screening for gastric cancer in Asia: current evidence and practice. *Lancet Oncol* 2008; 9: 279–287.
32. Smith JT, Pounder RE, Nwokolo CU, et al. Inappropriate hypergastrinaemia in asymptomatic healthy subjects infected with *Helicobacter pylori*. *Gut* 1990; 31: 522–525.
33. Li MY, Zhang DQ, Lu X, et al. Comparison of two serological methods in screening gastric cancer and its precancerous condition. *Zhonghua Nei Ke Za Zhi* 2018; 57: 907–911.
34. Yamaguchi Y, Nagata Y, Hiratsuka R, et al. Gastric cancer screening by combined assay for serum anti-*Helicobacter pylori* IgG antibody and serum pepsinogen levels–the ABC method. *Digestion* 2016; 93: 13–18.
35. Ni DQ, Lyu B, Bao HB, et al. Comparison of different serological methods in screening early gastric cancer. *Zhonghua Nei Ke Za Zhi* 2019; 58: 294–300.
36. Ghoshal UC, Kumar S, Krishnani N, et al. Serological assessment of gastric intestinal metaplasia and atrophy using pepsinogen-I, pepsinogen-II and gastrin-17 levels in a low incidence area of gastric cancer endemic for H. pylori infection. *Trop Gastroenterol* 2011; 32: 292–298.
37. Wang SM, Zheng RS, Zhang SW, et al. Epidemiological characteristics of gastric cancer in China, 2015. *Zhonghua Liu Xing Bing Xue Za Zhi* 2019; 40: 1517–1521.