International Journal of General Medicine Dovepress

Open Access Full Text Article ORIGINAL RESEARCH

Helicobacter Pylori Infection as the Predominant High-Risk Factor for Gastric Cancer Recurrence Post-Gastrectomy: An 8-Year Multicenter Retrospective Study

Yuan Liu[1,2,*], Xingchen Shang[1,*], Wenyi Du[1,*], Wei Shen[1], Yanfei Zhu[1]

[1]Department of General Surgery, Wuxi Medical Center of Nanjing Medical University, Wuxi, People's Republic of China; [2]Department of General Surgery, Tengzhou Central People's Hospital, Jining Medical College, Shandong, People's Republic of China

*These authors contributed equally to this work

Correspondence: Wei Shen; Yanfei Zhu, Department of General Surgery, Wuxi Medical Center of Nanjing Medical University, Wuxi, People's Republic of China, Tel +86133851 10723; +8617852061572, Email shenweiijs@outlook.com; wxsrmyy@outlook.com

Purpose: The reappearance of gastric cancer, a frequent postoperative complication following radical gastric cancer surgery, substantially impacts the near-term and far-reaching medical outlook of patients. The objective of this research was to create a machine learning algorithm that could recognize high-risk factors for gastric cancer recurrence and anticipate the correlation between gastric cancer recurrence and Helicobacter pylori (H. pylori) infection.

Patients and Methods: This investigation comprised 1234 patients diagnosed with gastric cancer, and 37 characteristic variables were obtained. Four machine learning algorithms, namely, extreme gradient boosting (XGBoost), random forest (RF), k-nearest neighbor algorithm (KNN), and multilayer perceptron (MLP), were implemented to develop the models. The k-fold cross-validation technique was utilized to perform internal validation of the four models, while independent datasets were employed for external validation of the models.

Results: In contrast to the other machine learning models, the XGBoost algorithm demonstrated superior predictive ability regarding high-risk factors for gastric cancer recurrence. The outcomes of Shapley additive explanation (SHAP) analysis revealed that tumor invasion depth, tumor lymph node metastasis, H. pylori infection, postoperative carcinoembryonic antigen (CEA), tumor size, and tumor number were risk elements for gastric cancer recurrence in patients, with H. pylori infection being the primary high-risk factor.

Conclusion: Out of the four machine learning models, the XGBoost algorithm exhibited superior performance in predicting the recurrence of gastric cancer. In addition, machine learning models can help clinicians identify key prognostic factors that are clinically meaningful for the application of personalized patient monitoring and immunotherapy.

Keywords: gastric tumor, gastrectomy, helicobacter pylori, immunotherapy, risk factor, machine learning

Introduction

Gastric cancer is among the most prevalent malignancies and the primary cause of cancer-related fatalities, especially in developing countries.[1,2] Global epidemiological surveys of tumors have demonstrated that the incidence of gastric cancer is progressively increasing owing to alterations in people's dietary habits.[3] Timely diagnosis and treatment of tumors are of utmost importance. In recent years, the implementation of comprehensive therapies such as immunotherapy and molecular targeted drugs has notably elevated the survival rate of patients with advanced gastric cancer.[4] Clinicians have also made efforts to eradicate gastric cancer using advanced surgical techniques to enhance the lifespan of patients.[5] Nonetheless, postoperative recurrence remains a significant challenge in the realm of oncology.[6] Approximately 10% of gastric cancer patients exhibit distant metastasis or tumor

Received: 3 July 2024
Accepted: 24 October 2024
Published: 29 October 2024

recurrence after postoperative review.[7] Helicobacter pylori (H. pylori) contributes significantly to the initial development of gastric cancer by inducing tumor gene expression and signaling pathways while altering the tumor microenvironment and disrupting the natural ecosystem between the tumor and host.[8,9] Nevertheless, the question of whether H. pylori has an impact on gastric cancer recurrence in postgastrectomy patients is also a highly debated issue.[10,11]

Presently, clinicians rely on a combination of H. pylori infection diagnosis, clinical presentation, and patient history to assess the link between H. pylori and gastric cancer recurrence.[12] However, this approach has limitations in terms of timeliness and subjectivity, thereby impeding accurate prediction. Imaging techniques such as computed tomography (CT) and magnetic resonance imaging (MRI), which are used in the diagnostic process, not only add to medical personnel workload but also escalate the financial burden on the patient's family. Additionally, some examination procedures are invasive and involve radiation exposure, causing potential harm to the patient. As a result, some researchers have opted to employ traditional regression models to predict the relationship between H. pylori and gastric cancer recurrence. Nonetheless, such an approach possesses limited discrimination and calibration capabilities.[13] With artificial intelligence making significant strides in the medical field, its principal branch - machine learning algorithms - can learn and analyze extensive data, discovering complex relationships and patterns between variables, thereby facilitating future disease occurrence prediction.[14] Compared to traditional prediction methods based on statistical methods and empirical rules, machine learning algorithms possess stronger adaptive and generalization capabilities. They can adapt to a wider and more complex range of data situations and avoid errors introduced by subjective factors of researchers and limitations of research methods. In this study, we analyzed the clinical information of patients with advanced gastric cancer and employed machine learning algorithms to establish a prediction model for tumor recurrence after gastrectomy. This will aid clinicians in formulating accurate and personalized treatment plans in a timely manner, thereby improving the postoperative survival quality of patients.

Materials and Methods

Study Subjects

In this study, we used data from the clinical databases of the Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi Second People's Hospital, and Shandong Provincial Hospital affiliated with Shandong First Medical University. The criteria for patient inclusion in this study were as follows: (1) the patient underwent either laparoscopic-assisted gastrectomy or conventional open gastrectomy; (2) the surgical team was composed of experienced senior doctors with independent experience in gastrectomy; and (3) the patient was diagnosed with gastric adenocarcinoma through postoperative pathology. Exclusion criteria for the cases were as follows: (1) patients with concomitant other malignancies; (2) patients diagnosed with distant metastasis of gastric cancer by pathological examination or imaging; (3) patients diagnosed with serious cardiovascular or respiratory diseases; (4) patients diagnosed with significant organ diseases such as liver and kidney; and (5) patients with missing cases, clinical data or visits. All patients in the study were followed up for at least 5 years after surgery. This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi Second People's Hospital, and Shandong Provincial Hospital affiliated with Shandong First Medical University, with approval number KY22085.

Diagnosis of Helicobacter Pylori Infection and Determination of Associated Factors

The diagnosis of H. pylori infection was based on three criteria: (1) postoperative bacterial culture of gastric mucosa, duodenal mucosa, gastric fluid, and expiratory samples to confirm the presence of positive H. pylori; (2) postoperative HE staining of gastric mucosal tissue sections to determine the presence of positive H. pylori; and (3) postoperative confirmation of H. pylori infection by urea breath test (UBT), fecal antigen test, and endoscopy active infection. The patient fulfilled all three criteria and was diagnosed with H. pylori infection.[15]

Study Design and Data Collection

The clinical data of patients were evaluated, encompassing various domains such as demographic characteristics, basic clinical features, medical history, preoperative as well as postoperative laboratory test indices, tumor characteristics, and intraoperative characteristics. Preoperative laboratory tests, including albumin (ALB), were conducted within 24 hours of surgery, while postoperative laboratory tests were conducted within 48 hours postoperatively and encompassed the H. pylori infection status, carcinoembryonic antigen (CEA), carbohydrate antigen 19–9 (CA19-9), carbohydrate antigen 72–4 (CA72-4), carbohydrate antigen 125 (CA125), neutrophil-to-lymphocyte ratio (NLR), procalcitonin (PCT), C-reactive protein (CRP), and serum amyloid A (SAA). Demographic information included sex, age, body mass index (BMI), and history of smoking and alcohol abuse. Basic clinical features comprised the American Society of Anesthesiologists physical status classification (ASA score), Nutrition Risk Screening 2002 (NRS2002) score, history of surgery, family history, history of adjuvant chemotherapy, and history of adjuvant radiotherapy. Medical history included anemia, diabetes mellitus, hypertension, hyperlipidemia, and coronary heart disease (CHD). Tumor characteristics included the tumor T-stage, N-stage, peripheral nerve invasion (PNI), tumor size, and tumor number. Intraoperative variables included the surgical approach, type of surgery, number of intraoperative lymph node dissections, anastomosis, type of anastomosis, and whether the surgery was performed as an emergency. The outcome indicator for this study was gastric cancer recurrence.

Statistical Analysis

Continuous variables are presented as medians and interquartile ranges (IQRs), while categorical variables are presented as frequencies and percentages. The chi-square test was used to compare differences between the two groups for categorical variables, and the *t* test was used for continuous variables that followed a normal distribution. For continuous variables that did not follow a normal distribution, the rank sum test was used. A two-tailed P value of <0.05 was considered statistically significant. All statistical analyses were performed using SPSS (version 30, International Business Machines Corporation, USA), R (version 4.4.1), and Python (version 3.10).

Development and Evaluation of Predictive Models for Machine Learning Algorithms

(1) Data preprocessing. Patients with gastric cancer who received treatment at Wuxi People's Hospital and Wuxi Second People's Hospital between January 2010 and January 2018 were selected as the internal validation group, while patients with gastric cancer who received treatment at the Provincial Hospital affiliated with the First Medical University of Shandong Province during the same period were selected as the external validation group. We employed the Random Sampling method for group allocation, dividing the entire internal validation set into two subsets: a training set (70%) and a test set (30%). The training set is utilized for model construction, while the test set serves to assess the model's performance. To ensure reproducibility in the random grouping process, we introduced a fixed random seed (random seed = 42). By consistently applying this random seed, we guarantee that the dataset is partitioned identically with each iteration, yielding the same training and test sets across all runs. (2) Data from the internal validation set underwent univariate analysis. Significant variables from the univariate analysis were subjected to logistic regression analysis to determine their independent influence on postoperative gastric cancer recurrence. Four models, namely, extreme gradient boosting (XGBoost), random forest (RF), multilayer perceptron (MLP), and k-nearest neighbor algorithm (KNN), were employed to score the importance of each independent influencing factor. The weight importance of each independent influencing factor was used to rank and score them. The top ten variables from the ranking of all four models were selected. In the XGBoost model, feature importance is assessed by calculating the gain each feature contributes to the splits. The importance score of each feature can be retrieved using the method get_booster().get_score(importance_type="weight"). In the Random Forest model, feature importance is evaluated by calculating the contribution of each feature to the information gain within the decision trees, with the relative importance extracted through the feature_importances_ attribute. For MLP, feature importance is typically derived from the network layer weights, and we employ Permutation Feature Importance to assess each feature's contribution. The KNN model does not inherently provide feature importance; instead, we utilize a feature selection method, such as recursive feature elimination, to evaluate

feature impact. (3) Build and evaluate prediction models. The selected feature variables were integrated into the prediction models of four machine learning algorithms, namely, XGBoost, MLP, RF, and KNN. To compare and select different model algorithms, k-fold cross-validation was used since it is easy to implement and has a lower bias evaluation capability compared to other methods. Hyperparameters were adjusted by grid search, and k-fold cross-validation was performed on the internal validation set using a resampling method with k=5. The dataset was divided into five groups, with one group used as a test dataset and the rest used as a training dataset. This process was repeated until each group had been used as a test dataset. Model evaluation metrics such as the area under the curve (AUC), accuracy, sensitivity, and specificity were calculated and averaged over the K-round fitness to derive the most accurate estimate of the model prediction performance. The models were evaluated for discrimination, calibration, and clinical utility, and the best model was selected for prediction analysis. Receiver operating characteristic (ROC) curves were used to determine the predictive efficacy of the model, calibration curves were used to assess agreement between the predicted and actual outcomes, and decision curve analysis (DCA) was used to evaluate the clinical utility of the model. The DCA curve starts at the intersection of the red curve with the All curve and ends at the intersection of the red curve with None, within which the corresponding patient can benefit. (4) External validation of the best model will be conducted using an external test set. ROC curves will be plotted to evaluate the predictive efficiency and generalizability of the model. (5) Model interpretation. The Shapley value, obtained through Shapley additive explanation (SHAP) analysis, allows us to determine the contribution of each feature in the sample to the prediction. Based on the Shapley values, two types of plots are constructed: the SHAP summary plot, which ranks the importance of risk factors, and the single-sample SHAP force plot, which analyzes and explains the prediction results of a single sample.

Results

Basic Clinical Information of the Patient

In this study, we included a total of 1234 patients diagnosed with gastric cancer, out of whom 117 (9.48%) patients experienced postoperative recurrence. The internal validation set comprised 877 gastric cancer patients, 82 (9.35%) of whom experienced recurrence. The external validation set comprised 357 gastric cancer patients, of whom 35 (9.8%) experienced recurrence (Figure 1A and B). The original data presented in the study are included in Table S1.

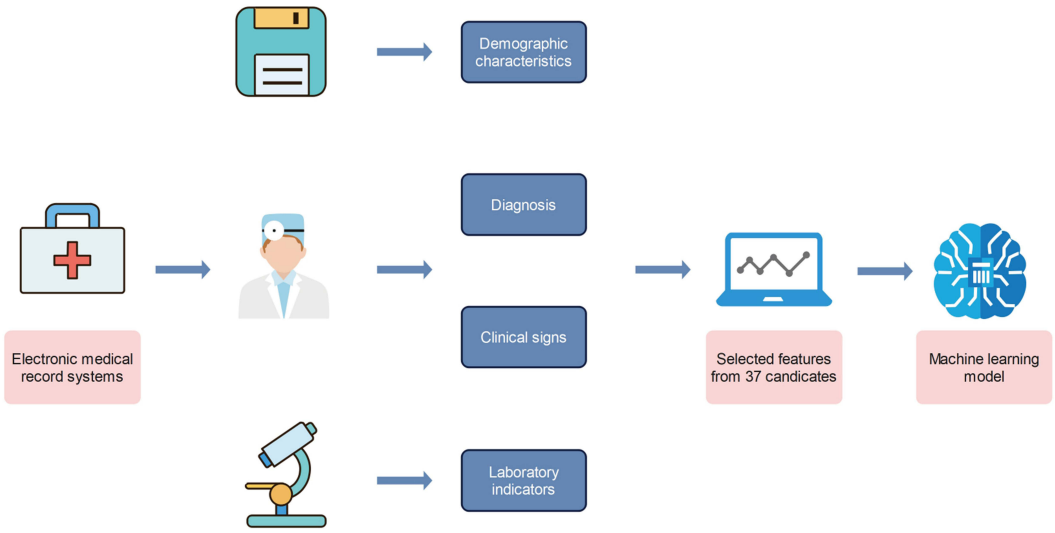
Screening for Risk Factors for Recurrence of Gastric Cancer

The findings from the univariate and multivariate analyses indicate that various factors, such as the depth of tumor invasion, tumor lymph node metastasis, tumor PNI, H. pylori infection, emergency surgery, postoperative CEA level, postoperative CA19-9 level, postoperative CA72-4 level, tumor size, and tumor number, have a significant impact on the postoperative recurrence of gastric cancer ($P<0.05$) (Table 1). The XGBoost, RF, MLP, and KNN models identified several risk factors, including tumors in T3 and T4 stages, tumors with lymph node metastasis, H. pylori infection, postoperative CEA levels $\geq$ 5 ng/mL, tumors with sizes $\geq$ 5 cm, and multiple tumors (Figure 2A–D). The variables selected for the predictive model in the combined analysis included tumor T stage, tumor with lymph node metastasis, H. pylori infection, postoperative CEA levels, tumor size, and number of tumors.

Model Building and Evaluation

The ROC curve analysis demonstrated that the RF model had the highest performance among the four models, with AUC values of 0.983 for the training set and 0.971 for the validation set (Table 2). The calibration curves of the models were consistent with the ideal curves, indicating good agreement between the predicted and observed outcomes. The DCA curves showed that all four models were associated with a net clinical benefit compared to a full or no treatment plan (Figure 3A–D). To evaluate the generalization ability of the models, a k-fold cross-validation method was employed (k=5). A sample size of 264 cases (30.10%) was randomly selected as the test set, while the remaining cases were used for training. The XGBoost model had the highest performance in both the validation set (AUC=0.9478±0.0298) and the test set (AUC=0.9695), with an accuracy of 0.9129 (Figure 4A–C). The RF model achieved an AUC of 0.8174±0.0887 in the validation set and an AUC of 0.9226 in the test set with an accuracy of 0.9091. The MLP model achieved an AUC of

(A)



(B)

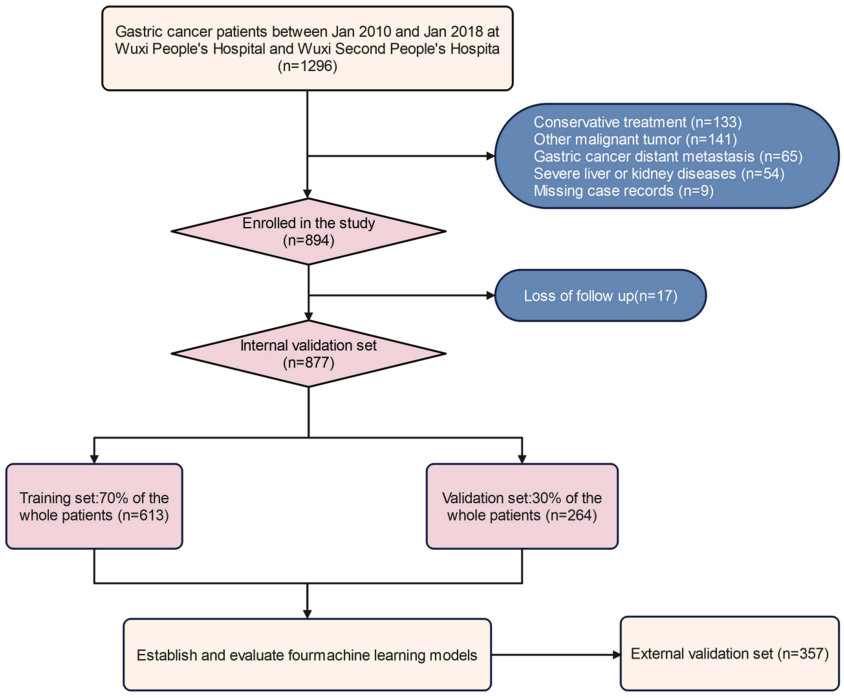


Figure 1 Model-making process and flowchart of the study. (**A**) Study design flow chart. (**B**) Flow diagram of patients included in the study.

Table 1 Univariate and Multivariate Analyses of Variables Related to Recurrence of Gastric Cancer

Variables		Univariate Analysis			Multivariate Analysis		
		OR	95% CI	P-value	OR	95% CI	P-value
Sex	Female	Reference					
	Male	1.253	[0.793,1.980]	0.334			
Age	<65	Reference			Reference		
	≥65	4.471	[2.803,7.132]	<0.001	1.713	[0.434,6.655]	0.434
BMI	<25 kg/m^2	Reference					
	≥25 kg/m^2	1.379	[0.822,2.313]	0.224			
ASA	<3	Reference					
	≥3	0.933	[0.592,1.470]	0.764			
ALB	≥30 g/l	Reference			Reference		
	<30 g/l	0.519	[0.310,0.869]	0.013	0.42	[0.091,1.616]	0.229
NRS2002 score	<3	Reference					
	≥3	0.751	[0.468,1.207]	0.237			
Family history	No	Reference					
	Yes	1.219	[0.586,2.538]	0.596			
Drinking history	No	Reference					
	Yes	1.336	[0.808,2.209]	0.259			
Smoking history	No	Reference					
	Yes	1.218	[0.727,2.039]	0.454			
Surgical history	No	Reference					
	Yes	1.167	[0.646,2.107]	0.609			
Anemia	No	Reference					
	Yes	1.203	[0.684,2.114]	0.521			
Hyperlipidemia	No	Reference					
	Yes	1.066	[0.591,1.921]	0.832			
Hypertension	No	Reference					
	Yes	0.658	[0.378,1.148]	0.141			
Diabetes	No	Reference					
	Yes	1.468	[0.897,2.403]	0.127			
CHD	No	Reference					
	Yes	1.487	[0.755,2.930]	0.251			
Adjuvant Radiotherapy	No	Reference			Reference		
	Yes	3.984	[2.415,6.572]	<0.001	2.67	[0.532,12.824]	0.22
Adjuvant Chemotherapy	No	Reference			Reference		
	Yes	6.156	[3.801,9.968]	<0.001	1.048	[0.308,3.381]	0.938
T-stage	T1~T2	Reference			Reference		
	T3~T4	10.565	[6.356,17.560]	<0.001	5.977	[1.814,21.992]	0.004
N-stage	N0	Reference			Reference		
	N1~N3	11.794	[7.199,19.322]	<0.001	16.302	[4.942,64.006]	<0.001
PNI	No	Reference			Reference		
	Yes	16.033	[9.279,27.703]	<0.001	5.261	[1.195,23.906]	0.028
Tumor number	<2	Reference			Reference		
	≥2	6.142	[3.821,9.872]	<0.001	10.404	[3.011,42.506]	<0.001
Tumor size	<5 cm	Reference			Reference		
	≥5 cm	14.713	[8.837,24.496]	<0.001	30.509	[7.026,162.378]	<0.001
Anastomosis method	Anastomosis instruments	Reference					
	Manual anastomosis	0.793	[0.501,1.253]	0.319			
Anastomosis type	Billroth I	Reference					
	Billroth II	0.927	[0.531,1.618]	0.791			
	Roux-en-Y	1.021	[0.591,1.765]	0.939			

(*Continued*)

Table 1 (Continued).

Variables		Univariate Analysis			Multivariate Analysis		
		OR	95% CI	P-value	OR	95% CI	P-value
Surgery type	Proximal gastrectomy	Reference					
	Distal gastrectomy	0.551	[0.302,1.008]	0.053			
	Total gastrectomy	0.853	[0.508,1.432]	0.547			
Surgical procedure	Laparoscopic surgery	Reference					
	Open surgery	0.727	[0.460,1.147]	0.171			
Emergency surgery	No	Reference			Reference		
	Yes	11.881	[7.181,19.658]	<0.001	16.939	[5.168,67.213]	<0.001
Lymph node dissection	<12	Reference					
	≥12	0.82	[0.486,1.384]	0.457			
HP infection	No	Reference			Reference		
	Yes	18.836	[10.881,32.606]	<0.001	10.721	[3.293,41.831]	<0.001
PCT level	<0.05 ng/mL	Reference					
	≥0.05 ng/mL	1.165	[0.610,2.224]	0.644			
CRP level	<10 mg/l	Reference			Reference		
	≥10 mg/l	2.286	[1.441,3.625]	<0.001	1.158	[0.33,4.009]	0.816
SAA level	<10 mg/l	Reference			Reference		
	≥10 mg/l	2.19	[1.282,3.742]	0.004	1.596	[0.349,6.913]	0.535
NLR	<3	Reference			Reference		
	≥3	5.863	[3.602,9.543]	<0.001	0.867	[0.209,3.302]	0.837
CEA level	<5 ng/mL	Reference			Reference		
	≥5 ng/mL	9.775	[5.968,16.011]	<0.001	9.784	[2.916,38.264]	<0.001
CA19-9 level	<37 U/mL	Reference			Reference		
	≥37 U/mL	9.227	[5.682,14.984]	<0.001	7.136	[2.338,24.449]	0.001
CA125 level	<35 U/mL	Reference			Reference		
	≥35 U/mL	4.112	[2.567,6.587]	<0.001	1.986	[0.565,7.144]	0.283
CA72-4 level	<7 U/mL	Reference			Reference		
	≥7 U/mL	7.64	[4.429,13.179]	<0.001	4.617	[1.236,18.075]	0.024

Abbreviations: OR, odds ratio; CI, confidence interval; BMI, body mass index; ASA, The American Society of Anesthesiologists; ALB, albumin; PCT, procalcitonin; CRP, C-reactive protein; SAA, serum amyloid A; NRS2002, nutrition risk screening 2002; CHD, coronary heart disease; PNI, peripheral nerve invasion; CEA, carcinoembryonic antigen, CA19-9, carbohydrate antigen 19–9, CA72-4, carbohydrate antigen 72–4, CA125, carbohydrate antigen 125, NLR, neutrophil-to-lymphocyte ratio.

0.8915±0.1205 in the validation set and an AUC of 0.8907 in the test set with an accuracy of 0.8788. The KNN model achieved an AUC of 0.9175±0.0534 in the validation set and an AUC of 0.8741 in the test set with an accuracy of 0.9508. Based on a comprehensive evaluation, the XGBoost algorithm was chosen to construct the prediction model in this study.

Model External Validation

The ROC curve analysis results demonstrated an AUC value of 0.82 for the external validation set, suggesting that the prediction model is highly precise in detecting the disease (Figure 4D).

Model Explanation

The SHAP summary plot revealed that the risk factors associated with postoperative recurrence of gastric cancer were ranked as H. pylori infection, tumors in T3 and T4 stages, tumors with a size of ≥5 cm, tumor lymph node metastasis, postoperative CEA levels of ≥5 ng/mL, and multiple tumors (Figure 5). The SHAP force diagram was used to predict the recurrence probability of four patients. For patient I, the model predicted a recurrence probability of 0.073, and the factors that increased this probability were tumor lymph node metastasis and H. pylori infection. For patient II, the model predicted a recurrence probability of 0.866,

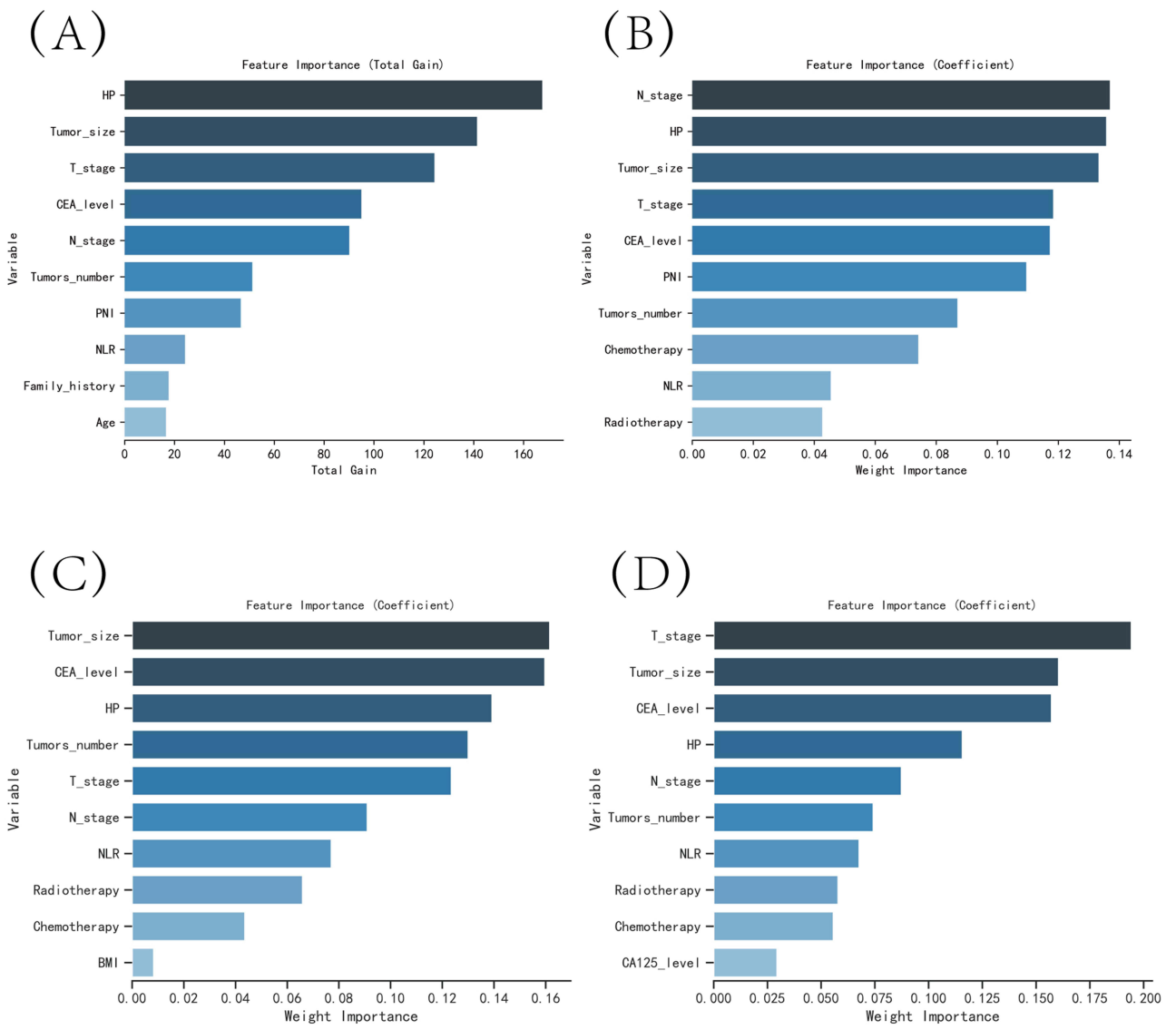


Figure 2 The variable influence factor ranking plots of the four models. (**A**) Variable importance ranking diagram of the XGBoost model. (**B**) Variable importance ranking diagram of the RF model. (**C**) Variable importance ranking diagram of the MLP model. (**D**) Variable importance ranking diagram of the KNN model.

and the factors that increased the probability were T3 and T4 tumors, tumor lymph node metastasis, postoperative CEA levels of ≥5 ng/mL, and H. pylori infection. For patient III, the model predicted a recurrence probability of 0.062, and the factors that increased this probability were tumor lymph node metastasis and tumor size ≥5 cm. For patient IV, the model predicted

Table 2 Evaluation of the Performance of the Four Models

		AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
KNN	Training set	0.963 (0.931–0.995)	0.963 (0.957–0.968)	0.936 (0.890–0.982)	0.922 (0.898–0.946)
	Validation set	0.898 (0.776–0.999)	0.954 (0.936–0.973)	0.828 (0.754–0.903)	0.929 (0.888–0.970)
XGBoost	Training set	0.983 (0.967–0.998)	0.937 (0.916–0.957)	0.952 (0.926–0.978)	0.923 (0.898–0.948)
	Validation set	0.971 (0.935–1.000)	0.914 (0.889–0.939)	0.965 (0.921–1.000)	0.898 (0.819–0.976)
RF	Training set	0.838 (0.780–0.897)	0.900 (0.859–0.942)	0.786 (0.610–0.961)	0.824 (0.755–0.894)
	Validation set	0.862 (0.744–0.976)	0.909 (0.879–0.939)	0.830 (0.714–0.945)	0.838 (0.775–0.901)
MLP	Training set	0.932 (0.880–0.984)	0.909 (0.886–0.932)	0.909 (0.866–0.952)	0.906 (0.874–0.938)
	Validation set	0.807 (0.681–0.918)	0.889 (0.875–0.904)	0.871 (0.743–0.999)	0.804 (0.609–0.999)

Abbreviations: CI, confidence interval; KNN, k-nearest neighbor; XGBoost, extreme gradient boosting; RF, random forest; MLP, multilayer perceptron; AUC, area under the curve.

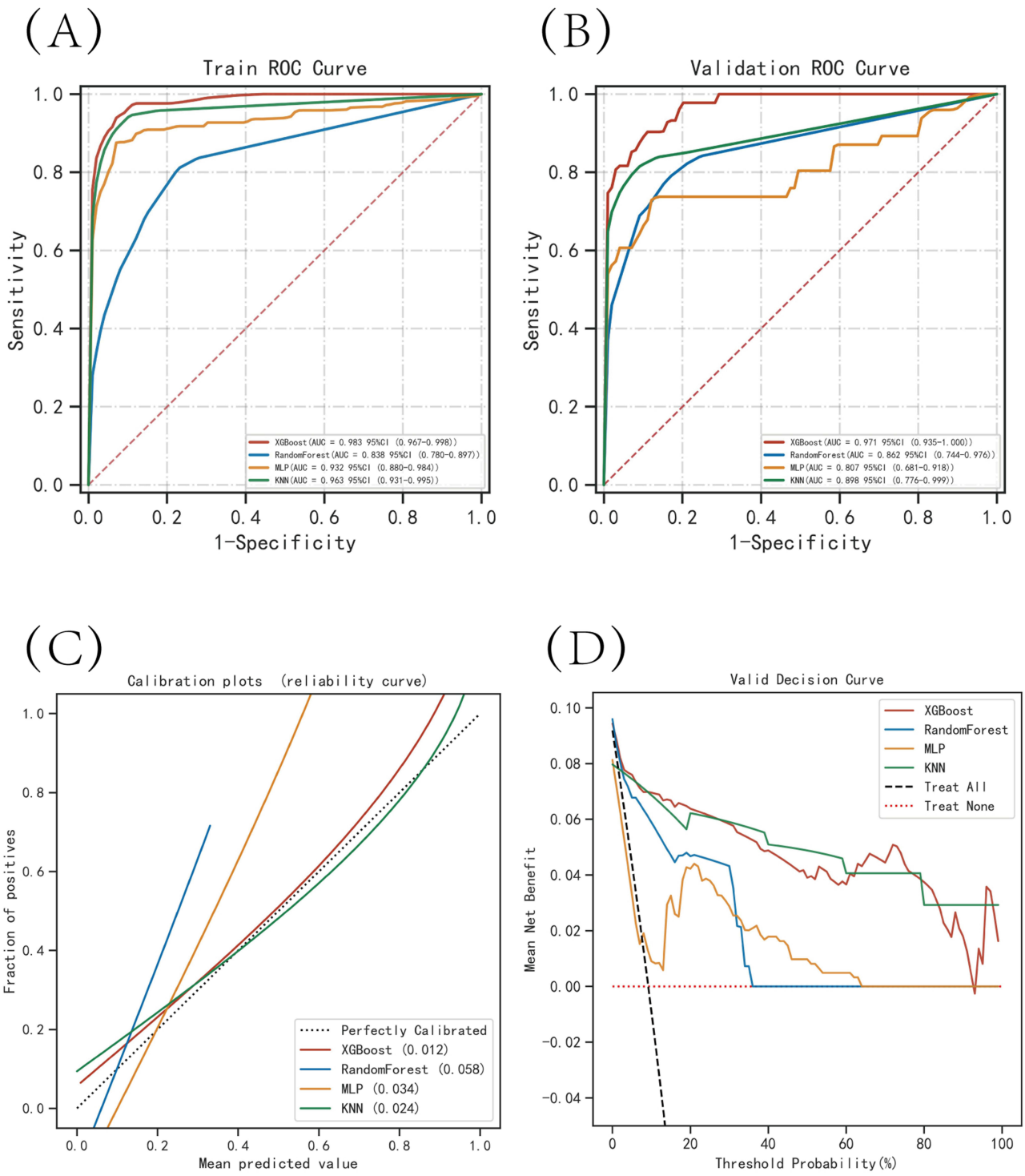


Figure 3 Evaluation of the four models for predicting recurrence of gastric cancer. (**A**) ROC curves for the training set of the four models. (**B**) ROC curves for the validation set of the four models. (**C**) Calibration plots of the four models. The 45° dotted line on each graph represents the perfect match between the observed (y-axis) and predicted (x-axis) complication probabilities. A closer distance between two curves indicates greater accuracy. (**D**) DCA curves of the four models. The intersection of the red curve and the All curve is the starting point, and the intersection of the red curve and the None curve is the node within which the corresponding patients can benefit.

a recurrence probability of 0.362, and the factors that increased this probability were tumor size ≥5 cm and H. pylori infection (Figure 6A–D).

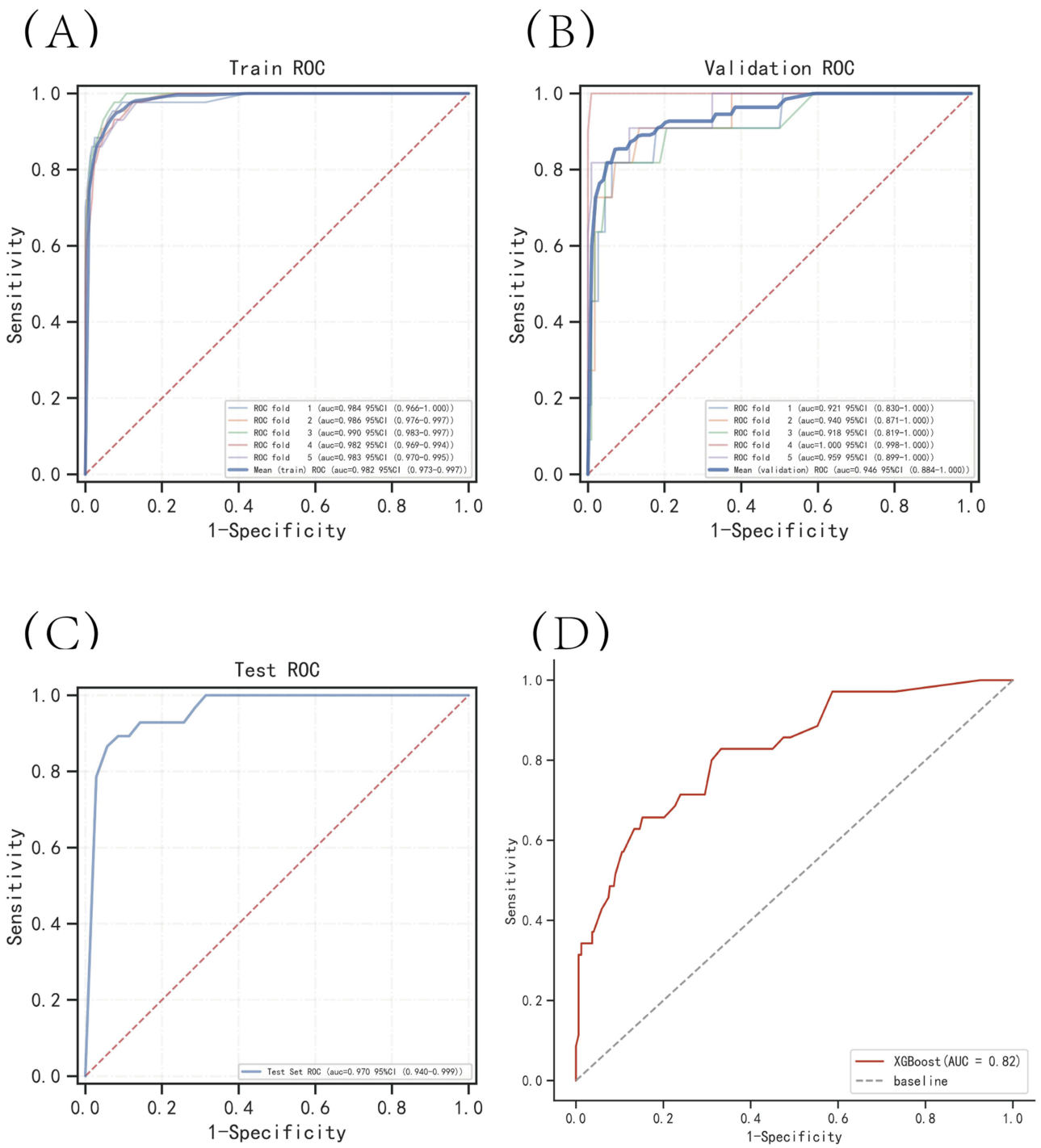


Figure 4 Internal validation of the XGBoost model. (**A**) ROC curve of the XGBoost model for the training set. (**B**) ROC curve of the XGBoost model for the validation set. (**C**) ROC curve of the XGBoost model for the test set. (**D**) External validation of the XGBoost model.

Discussion

This investigation appraised the risk prediction models developed by four machine learning algorithms, with the XGBoost algorithm displaying the highest accuracy.[16] In contrast to the RF algorithm, XGBoost employs an adaptive gradient boosting algorithm that can automatically select the optimal splitting point and tree depth, which, in turn, enhances the prediction performance. Furthermore, XGBoost fully takes into account the regularization issue and can effectively prevent model overfitting.[17] Although the KNN algorithm has greater accuracy and can efficiently avoid the

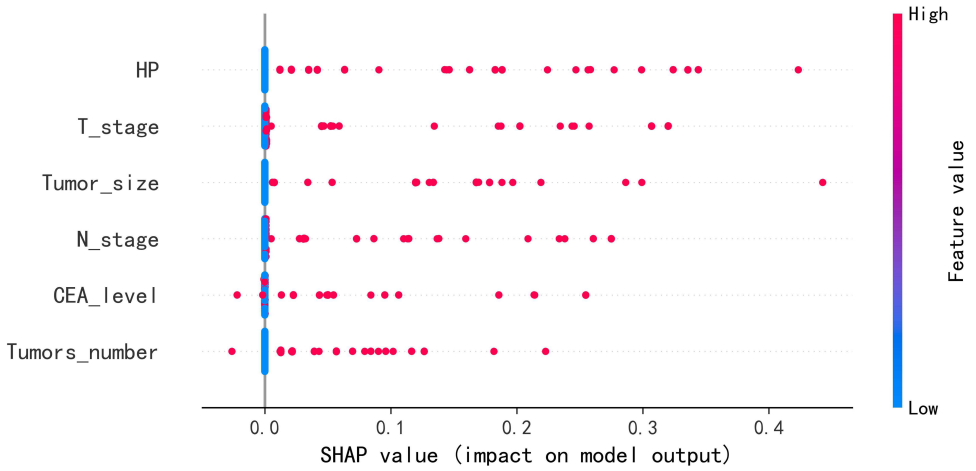


Figure 5 SHAP summary plot. Risk factors are arranged along the y-axis based on their importance, which is given by the mean of their absolute Shapley values. The higher the risk factor is positioned in the plot, the more important it is for the model.

overfitting problem, such algorithms need to search for K nearest neighbors in the training set for each test sample and calculate their distances for classification or regression prediction. This is associated with high computational complexity, and the algorithms are less stable and slower when addressing issues with multiple features and large samples.[18] The XGBoost algorithm is better suited to multidimensional investigations and reduces the computational effort as well as the training time. Most importantly, XGBoost offers a feature importance assessment function that helps users better comprehend the contribution of features in the dataset to the prediction results, thereby enhancing the interpretability of the algorithm. Therefore, after conducting an all-inclusive comparison of the four machine learning algorithms, this investigation chose to utilize the XGBoost algorithm to construct a model for forecasting postoperative recurrence in gastric cancer patients.

In clinical research, the relationship between various risk factors and patient prognosis is often nonlinear, particularly in cancer research. This means that traditional models may not adequately fit the data or make accurate predictions. Machine learning, however, is capable of identifying and understanding complex patterns and nonlinear relationships, making it a potentially superior approach in medical research. DeGregory et al[19] have already validated the effectiveness of machine learning algorithms for clinical diagnosis and prognosis, showing that this artificial intelligence technique can accurately predict adverse outcomes in disease progression. Machine learning algorithms have also played a significant role in building the predictive model for the current study. This model can assist clinical decision makers in identifying high-risk patients more accurately and promptly initiating appropriate interventions to improve patient prognosis. Moreover, the model can aid medical institutions in rationally allocating medical resources, monitoring vital signs of high-risk patients, and ultimately enhancing the survival rate of gastric cancer patients.

The current investigation employed SHAP analysis to rank the risk factors for gastric cancer recurrence, and H. pylori infection was found to be the foremost among all high-risk factors. We posit that H. pylori infection facilitates the growth and survival of cancer cells by perturbing the normal interaction between the tumor and the host.[20] The primary mechanisms of action involve reducing the abundance of beneficial bacteria, such as lactobacilli and bifidobacteria, which impairs the inflammatory environment and thus promotes the development of gastric cancer.[21] Additionally, H. pylori hinders the activity of T cells and natural killer cells, fosters the recruitment of immunosuppressive cells, and influences the immune response in the stomach, thereby hampering the immune surveillance and clearance function of the body.[22] Furthermore, Helicobacter pylori can foster the growth and survival of cancer cells by eliciting the production of cytokines, such as interleukin-1β, tumor necrosis factor-alpha, and interleukin-6, thus creating a proinflammatory microenvironment.[23] From a genetic perspective, the infection gives rise to the generation of reactive oxygen and nitrogen species, augmenting the risk of cancer development. Li established a robust correlation between H. pylori and the activation of oncogenes, including c-Met and β-catenin, as well as the inactivation of tumor suppressor genes, such as

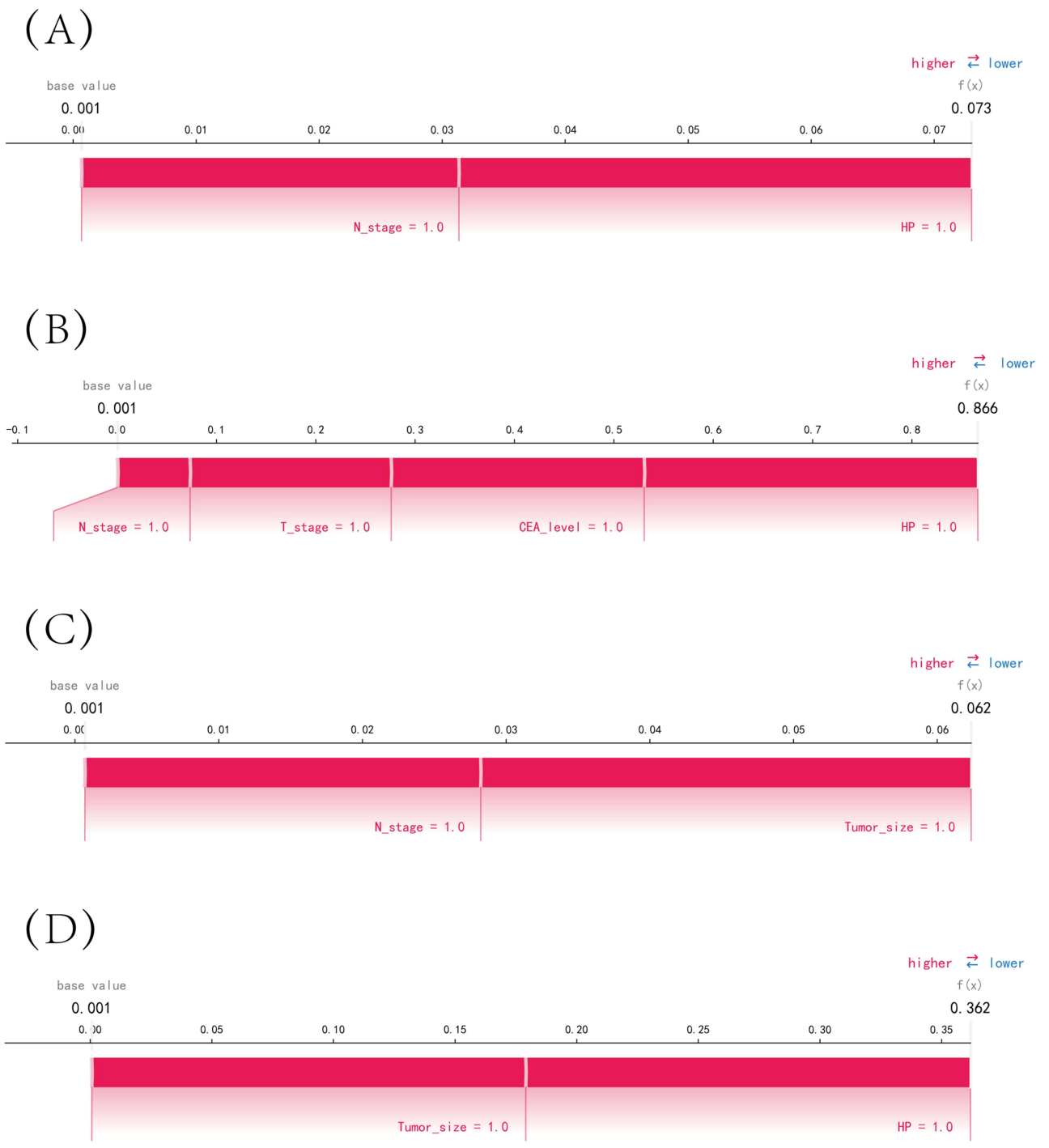


Figure 6 SHAP force plot. The contributing variables are arranged in the horizontal line, sorted by the absolute value of their impact. Blue represents features that have a negative effect on disease prediction, with a decrease in SHAP values; red represents features that have a positive effect on disease prediction, with an increase in SHAP values. (**A**) Predictive Analysis of Patient I. (**B**) Predictive Analysis of Patient II. (**C**) Predictive Analysis of Patient III. (**D**) Predictive Analysis of Patient IV.

p53 and E-cadherin.[24] These findings lend further support to the results of the present study. This study evaluated the utility of four distinct samples to assess the viability of a model for predicting gastric cancer recurrence. Analysis of Sample II identified H. pylori infection as a prominent risk factor, a finding we believe is intimately tied to its role in promoting tumor angiogenesis. The intricate anatomy of the stomach poses significant challenges for surgeons to accurately gauge the extent of tumor invasion during surgery. Additionally, conducting rapid intraoperative pathology

to ensure negative margins is often impractical, which increases the probability that residual tumor vasculature may serve as a source of recurrence. This risk is particularly heightened in H. pylori-infected patients, as the bacterium enhances the expression of vascular endothelial growth factor (VEGF), thereby fostering neovascularization within the tumor microenvironment, leading to enhanced tumor proliferation and migration.[25] Following the eradication of H. pylori, VEGF expression diminishes, resulting in a reduction in tumor angiogenesis, which in turn restricts the nutrient supply to tumor cells and curtails their growth and proliferation. Furthermore, H. pylori upregulates the expression of matrix metalloproteinase-9 (MMP-9), an enzyme that degrades the extracellular matrix, thus increasing the risk of tumor metastasis and recurrence. Eradication of H. pylori, however, plays a pivotal role in restraining the invasive properties of tumor cells.[26] Additionally, we posit that in patients with chronic atrophic gastritis and intestinal epithelial hyperplasia, H. pylori eradication leads to varying degrees of gastric mucosal repair over months to years, reversing or inhibiting the progression of precancerous lesions, thereby preventing gastric cancer formation. Early eradication of H. pylori in patients without postoperative gastric cancer precludes their entry into the Correa cascade. In these individuals, the gastric mucosa is less exposed to carcinogenic stimuli, reducing the incidence of lesions such as atypical hyperplasia and substantially lowering the risk of gastric cancer development.[27,28]

Similar to previous investigations, the current study demonstrates a correlation between the depth of tumor infiltration into the tissue, lymphatic metastasis, and an increased risk of gastric cancer recurrence postsurgery.[6] Malignant and biologically active tumor cells are capable of degrading the extracellular matrix and basement membrane with various protein hydrolases, leading to detachment from the origin site. The detached tumor cells infiltrate the surrounding normal tissues and enter the adjacent lymph nodes. Due to the rich blood vessels in the perigastric omentum, gastric cancer is prone to vascular invasion, causing tumor cells to return to the liver via the portal vein system, ultimately leading to postoperative recurrence or metastasis.[29] Furthermore, the tumors may metastasize to retroperitoneal organs via lymph nodes, with subtle clinical manifestations and difficulty in diagnosis by imaging examinations, thereby increasing the risk of postoperative tumor recurrence.[30,31] The findings from David's study also suggest that lymph node metastasis is closely associated with tumor recurrence, while Radespiel observed a higher mortality rate and an increased likelihood of tumor recurrence with a greater number of lymph node metastases.[32] Thus, it is crucial for surgeons to thoroughly remove relevant lymph nodes during radical surgery for gastric cancer while avoiding tumor dissemination to the abdominal cavity.

Similarly, larger tumors are associated with poor prognosis in patients. We hypothesize that larger tumor cells have a higher proliferation rate and thus generate more tumor vessels. Tomisaki examined 175 patients with gastrointestinal tumors and found a strong correlation between metastatic recurrence of tumors and tumor microvessel density (MVD). The higher the tumor microvessel density, the greater the risk of tumor cells entering the circulatory system.[33] Additionally, Park demonstrated that tumor cells from larger tumors have a higher likelihood of shedding and entering the abdominal and pelvic cavities as well as vascular tissues, thereby increasing the potential risk of tumor recurrence after surgery.[34] The occurrence of tumor recurrence is equally likely in patients with multiple gastric cancers. Surgical removal of the primary tumor decreases the concentration of tumor growth inhibitory factors and accelerates residual tumor recurrence. Li et al tested this hypothesis using two groups of mouse models. The experimental group of mice underwent conventional tumor resection, while the control group underwent sham surgery. They found that there were significant differences in the size of tumor growth and the degree of recurrence in the experimental group compared to the control group.[35]

The findings of the current investigation indicate that patients with higher CEA levels following radical surgery for gastric cancer are more susceptible to experiencing gastric cancer recurrence. CEA, considered by Gold to be an acidic glycoprotein expressed by normal human mucosal cells, was previously regarded as nonspecific for the diagnosis of gastrointestinal tumors.[36,37] However, over the past decade, with progress in medical testing techniques, medical practitioners have gradually recognized the value of CEA. Polat conducted a prospective analysis to examine the association between serum levels of tumor markers and clinical variables in patients with gastrointestinal tumors. Another study by Tsuyoshi et al demonstrated that for most patients, serum CEA levels reverted to normal three months after radical tumor therapy, while a proportion of patients whose postoperative CEA levels did not change from preoperative levels had rapid tumor recurrence. An increase in postoperative CEA levels can be considered a marker

for gastric cancer recurrence, which is consistent with the current investigation's results.[38,39] Recently, some medical practitioners have used a combination of preoperative CEA, CA19-9, cytokeratin-1 (CK-1), and mucin-1 (MUC-1) to identify gastrointestinal tumors in patients. This approach can enhance the sensitivity and specificity of tumor surveillance while also assessing tumor stage and metastasis and aiding in predicting patients' postoperative recurrence.[40]

The present investigation also included surgical modality as a factor to evaluate tumor recurrence and found no significant difference between the two surgical modalities. However, this topic is still controversial in clinical practice. Aasmund[41] argued that laparoscopic surgery, which adopts a minimally invasive approach, has a lower impact on the patient's immune system and a lower likelihood of tumor recurrence in postoperative patients. In contrast, Mirow[42] suggested that the use of trocars in laparoscopic surgery may result in the development of tumor implantation. Thus, it is recommended that medical practitioners choose minimally invasive surgical approaches as much as possible to minimize trauma while treating patients with colon cancer. The surgeon must strictly adhere to the principle of operating on a tumor-free area and avoid contacting the tumor when placing the trocar to decrease the risk of tumor dissemination.

Limitations

The present investigation conducted a comprehensive evaluation of the model in terms of discrimination, calibration, and clinical utility. However, the study has certain limitations. Although the study encompassed multiple aspects of risk factors, it did not take into account imaging factors. Furthermore, even though the machine learning algorithms were more accurate, their models were more complex and less interpretable The entire computational and decision-making process of the model operates as a black box, which is not as intuitive and transparent as the logistic regression model.[43] Moreover, the design of this study did not explicitly distinguish between past and present H. pylori infections. The study employed current diagnostic methods widely used in clinical practice, and the infection status of all patients was determined based on active infections confirmed through intraoperative and postoperative test results, without assessing any history of prior H. pylori infection. Future research could investigate the potential influence of previous H. pylori infection on postoperative recurrence by designing study models that specifically differentiate between historical and active infections. In our study, H. pylori was successfully isolated; however, the specific anatomical region (corpus or antrum) from which the samples were taken was not consistently recorded. We recognize that differences in H. pylori colonization across various stomach regions (eg, corpus versus antrum) may influence clinical outcomes. Unfortunately, due to the retrospective nature of our data collection, this detailed information was not consistently available for all patients. Consequently, further large-scale, multicenter, international studies are required to validate the robustness of our findings.

Conclusion

The current investigation involved the development of a model based on the XGBoost machine learning algorithm to predict the risk of tumor recurrence following radical surgery for gastric cancer. The model demonstrated excellent predictive accuracy and clinical utility, thus enabling surgeons to diagnose patients promptly. The model highlighted that gastric cancer recurrence was significantly associated with various factors, such as H. pylori infection, T3 and T4 tumors, tumors larger than or equal to 5 cm, tumor lymph node metastasis, postoperative CEA levels equal to or higher than 5 ng/mL, and multiple tumors.

Data Sharing Statement

The original data presented in the study are included in the Raw Data/Table S1, and further inquiries can be directed to the corresponding author (shenweiijs@outlook.com).

Ethics Statement

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Wuxi People's Hospital, with approval number KY22085. The review committee waived the requirement for written informed consent because of the retrospective nature of the study. Prior to analysis, confidential patient information was deleted from the entire data set.

Funding

This work was supported by the Top Talent Support Program for young and middle-aged people of Wuxi Health Committee (Grant No. HB2020007).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Cao M, Li H, Sun D, et al. Current cancer burden in China: epidemiology, etiology, and prevention. *Cancer Biol Med*. 2022;19(8):1121–1138. doi:10.20892/j.issn.2095-3941.2022.0231
2. Sekiguchi M, Oda I, Matsuda T, Saito Y. Epidemiological trends and future perspectives of gastric cancer in Eastern Asia. *Digestion*. 2022;103(1):22–28. doi:10.1159/000518483
3. Morgan E, Arnold M, Camargo MC, et al. The current and future incidence and mortality of gastric cancer in 185 countries, 2020-40: a population-based modelling study. *EClinicalMedicine*. 2022;47:101404. doi:10.1016/j.eclinm.2022.101404
4. Li K, Zhang A, Li X, Zhang H, Zhao L. Advances in clinical immunotherapy for gastric cancer. *Biochim Biophys Acta Rev Cancer*. 2021;1876(2):188615. doi:10.1016/j.bbcan.2021.188615
5. Kakinuma D, Arai H, Yasuda T, et al. Treatment of gastric cancer in Japan. *J Nippon Med Sch*. 2021;88(3):156–162. doi:10.1272/jnms.JNMS.2021_88-315
6. Xiang L, Jin S, Zheng P, et al. Risk assessment and preventive treatment for peritoneal recurrence following radical resection for gastric cancer. *Front Oncol*. 2021;11:778152. doi:10.3389/fonc.2021.778152
7. Shibata C, Nakano T, Yasumoto A, et al. Comparison of cea and ca19-9 as a predictive factor for recurrence after curative gastrectomy in gastric cancer. *BMC Surg*. 2022;22(1):213. doi:10.1186/s12893-022-01667-z
8. Alipour M. Molecular mechanism of helicobacter pylori-induced gastric cancer. *J Gastrointest Cancer*. 2021;52(1):23–30. doi:10.1007/s12029-020-00518-5
9. Liatsos C, Papaefthymiou A, Kyriakos N, et al. Helicobacter pylori, gastric microbiota and gastric cancer relationship: unrolling the Tangle. *World J Gastrointest Oncol*. 2022;14(5):959–972. doi:10.4251/wjgo.v14.i5.959
10. Abu Abeid A, Abeid SA, Nizri E, Kuriansky J, Lahat G, Dayan D. The association of helicobacter pylori, eradication, and early complications of laparoscopic sleeve gastrectomy. *Obes Surg*. 2022;32(5):1617–1623. doi:10.1007/s11695-022-05996-z
11. Malfertheiner P, Fry LC, Mönkemüller K. Can gastric cancer be prevented by helicobacter pylori eradication? *Best Pract Res Clin Gastroenterol*. 2006;20(4):709–719. doi:10.1016/j.bpg.2006.04.005
12. Şen O, Oray Ş, Gökhan Türkçapar A. Is there a relationship between helicobacter pylori and gerd before laparoscopic sleeve gastrectomy? *Surg Laparosc Endosc Percutan Tech*. 2022;32(6):692–695. doi:10.1097/sle.0000000000001122
13. Turan G, Kocaöz S. Helicobacter pylori infection prevalence and histopathologic findings in laparoscopic sleeve gastrectomy. *Obes Surg*. 2019;29(11):3674–3679. doi:10.1007/s11695-019-04052-7
14. Niu PH, Zhao LL, Wu HL, Zhao DB, Chen YT. Artificial intelligence in gastric cancer: application and future perspectives. *World J Gastroenterol*. 2020;26(36):5408–5419. doi:10.3748/wjg.v26.i36.5408
15. Wang YK, Kuo FC, Liu CJ, et al. Diagnosis of helicobacter pylori infection: current options and developments. *World J Gastroenterol*. 2015;21(40):11221–11235. doi:10.3748/wjg.v21.i40.11221
16. Hou N, Li M, He L, et al. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. *J Transl Med*. 2020;18(1):462. doi:10.1186/s12967-020-02620-5
17. Tseng PY, Chen YT, Wang CH, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care*. 2020;24(1):478. doi:10.1186/s13054-020-03179-9
18. Zhou X, Wang H, Xu C, et al. Application of Knn and Svm to predict the prognosis of advanced schistosomiasis. *Parasitol Res*. 2022;121(8):2457–2460. doi:10.1007/s00436-022-07583-8
19. DeGregory KW, Kuiper P, DeSilvio T, et al. A review of machine learning in obesity. *Obes Rev*. 2018;19(5):668–685. doi:10.1111/obr.12667
20. Baj J, Korona-Głowniak I, Forma A, et al. Mechanisms of the epithelial-mesenchymal transition and tumor microenvironment in helicobacter pylori-induced gastric cancer. *Cells*. 2020;9(4). doi:10.3390/cells9041055
21. Zuo F, Somiah T, Gebremariam HG, Jonsson AB. Lactobacilli downregulate transcription factors in helicobacter pylori that affect motility, acid tolerance and antimicrobial peptide survival. *Int J Mol Sci*. 2022;23(24). doi:10.3390/ijms232415451
22. Sipos G, Altdorfer K, Pongor E, Chen LP, Fehér E. Neuroimmune link in the mucosa of chronic gastritis with helicobacter pylori infection. *Dig Dis Sci*. 2006;51(10):1810–1817. doi:10.1007/s10620-006-9085-5
23. Figueiredo CA, Marques CR, Costa Rdos S, da Silva HB, Alcantara-Neves NM. Cytokines, cytokine gene polymorphisms and helicobacter pylori infection: friend or foe? *World J Gastroenterol*. 2014;20(18):5235–5243. doi:10.3748/wjg.v20.i18.5235
24. Li N, Tang B, Jia YP, et al. Helicobacter pylori caga protein negatively regulates autophagy and promotes inflammatory response Via C-Met-Pi3k /Akt-mtor signaling pathway. *Front Cell Infect Microbiol*. 2017;7:417. doi:10.3389/fcimb.2017.00417
25. Siregar G, Sari D, Sungkar T. Serum VEGF levels in helicobacter pylori infection and correlation with helicobacter pylori caga and vaca genes. *Open Access Maced J Med Sci*. 2017;5(2):137–141. doi:10.3889/oamjms.2017.031
26. Bagheri N, Sadeghiani M, Rahimian G, et al. Correlation between expression of mmp-9 and mmp-3 in helicobacter pylori infected patients with different gastroduodenal diseases. *Arab J Gastroenterol*. 2018;19(4):148–154. doi:10.1016/j.ajg.2018.11.001
27. Deng R, Zheng H, Cai H, Li M, Shi Y, Ding S. Effects of helicobacter pylori on tumor microenvironment and immunotherapy responses. *Front Immunol*. 2022;13:923477. doi:10.3389/fimmu.2022.923477

28. Oster P, Vaillant L, Riva E, et al. Helicobacter pylori infection has a detrimental impact on the efficacy of cancer immunotherapies. *Gut*. 2022;71(3):457–466. doi:10.1136/gutjnl-2020-323392
29. Zhou H, Liu Z, Wang Y, et al. Colorectal liver metastasis: molecular mechanism and interventional therapy. *Signal Transduct Target Ther*. 2022;7(1):70. doi:10.1038/s41392-022-00922-2
30. Shibata D, Paty PB, Guillem JG, Wong WD, Cohen AM. Surgical management of isolated retroperitoneal recurrences of colorectal carcinoma. *Dis Colon Rectum*. 2002;45(6):795–801. doi:10.1007/s10350-004-6300-3
31. Larsen SG, Wiig JN, Giercksky KE. Hydronephrosis as a prognostic factor in pelvic recurrence from rectal and colon carcinomas. *Am J Surg*. 2005;190(1):55–60. doi:10.1016/j.amjsurg.2004.07.043
32. Radespiel-Tröger M, Hohenberger W, Reingruber B. Improved prediction of recurrence after curative resection of colon carcinoma using tree-based risk stratification. *Cancer*. 2004;100(5):958–967. doi:10.1002/cncr.20065
33. Tomisaki S, Ohno S, Ichiyoshi Y, Kuwano H, Maehara Y, Sugimachi K. Microvessel quantification and its possible relation with liver metastasis in colorectal cancer. *Cancer*. 1996;77(8 Suppl):1722–1728. doi:10.1002/(sici)1097-0142(19960415)77:8<1722::Aid-cncr46>3.0.Co;2-z
34. Park YJ, Park KJ, Park JG, Lee KU, Choe KJ, Kim JP. Prognostic factors in 2230 Korean colorectal cancer patients: analysis of consecutively operated cases. *World J Surg*. 1999;23(7):721–726. doi:10.1007/pl00012376
35. Li TS, Kaneda Y, Ueda K, Hamano K, Zempo N, Esato K. The influence of tumour resection on angiostatin levels and tumour growth--an experimental study in tumour-bearing mice. *Eur J Cancer*. 2001;37(17):2283–2288. doi:10.1016/s0959-8049(01)00281-7
36. Gold P, Freedman SO. Specific carcinoembryonic antigens of the human digestive system. *J Exp Med*. 1965;122(3):467–481. doi:10.1084/jem.122.3.467
37. Seth J, Sturgeon CM, Hanning I. Carcinoembryonic antigen. *Lancet*. 1988;1(8599):1399. doi:10.1016/s0140-6736(88)92215-5
38. Polat E, Duman U, Duman M, et al. Diagnostic value of preoperative serum carcinoembryonic antigen and carbohydrate antigen 19-9 in colorectal cancer. *Curr Oncol*. 2014;21(1):e1–7. doi:10.3747/co.21.1711
39. Konishi T, Shimada Y, Hsu M, et al. Association of preoperative and postoperative serum carcinoembryonic antigen and colon cancer outcome. *JAMA Oncol*. 2018;4(3):309–315. doi:10.1001/jamaoncol.2017.4420
40. Attallah AM, El-Far M, Ibrahim AR, et al. Clinical value of a diagnostic score for colon cancer based on serum cea, ca19-9, cytokeratin-1 and mucin-1. *Br J Biomed Sci*. 2018;75(3):122–127. doi:10.1080/09674845.2018.1456309
41. Fretland AA, Sokolov A, Postriganova N, et al. Inflammatory response after laparoscopic versus open resection of colorectal liver metastases: data from the Oslo-comet trial. *Medicine*. 2015;94(42):e1786. doi:10.1097/md.0000000000001786
42. Mirow L. Trocar site recurrence in laparoscopic surgery for colorectal cancer. *Tech Coloproctol*. 2002;6(3):197–198. doi:10.1007/s101510200044
43. Pua YH, Kang H, Thumboo J, et al. Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc*. 2020;28(10):3207–3216. doi:10.1007/s00167-019-05822-7