94450_Auto_Edited-check.docx
Retrospective Study

Establishing and clinically validating a machine learning model for predicting unplanned reoperation risk in colorectal cancer

Cai LQ et al. Predicting colorectal surgery reoperations

Li-Qun Cai, Da-Qing Yang, Rong-Jian Wang, He Huang, Yi-Xiong Shi

Abstract

BACKGROUND

This study aimed to establish and validate a risk prediction model of postoperative unplanned reoperation in colorectal cancer.

AIM

To develop and validate a machine learning model for predicting unplanned reoperation risk in colorectal cancer patients.

METHODS

Data of patients treated for colorectal cancer (n = 2044) at the First Affiliated Hospital of Wenzhou Medical University and Wenzhou Central Hospital from March 2020 to March 2022 were retrospectively collected. Patients were divided into an experimental group (n = 60) and a control group (n = 1984) according to unplanned reoperation occurrence. Patients were also divided into a training group and a validation group (7:3 ratio). We used three different machine learning methods to screen characteristic variables. A nomogram was created based on multifactor logistic regression, and the
model performance was assessed using receiver operating characteristic curve, calibration curve, Hosmer-Lemeshow test, and decision curve analysis. The risk scores of the two groups were calculated and compared to validate the model.

RESULTS
More patients in the experimental group were ≥ 60 years old, male, and had a history of hypertension, laparotomy, and hypoproteinemia, compared to the control group. Multiple logistic regression analysis confirmed the following as independent risk factors for unplanned reoperation ($P < 0.05$): Prognostic Nutritional Index value, history of laparotomy, hypertension, or stroke, hypoproteinemia, age, tumor-node-metastasis staging, surgical time, gender, and American Society of Anesthesiologists classification. Receiver operating characteristic curve analysis showed that the model had good discrimination and clinical utility.

CONCLUSION
This study used a machine learning approach to build a model that accurately predicts the risk of postoperative unplanned reoperation in patients with colorectal cancer, which can improve treatment decisions and prognosis.

Key Words: Colorectal cancer; Postoperative unplanned reoperation; Unplanned reoperation; Clinical validation; Nomogram; Machine learning models


Core Tip: This study developed a machine learning model to predict unplanned reoperations in colorectal cancer patients, using data from two hospitals over two years. It employed support vector machine, least absolute shrinkage and selection operator,
and extreme gradient boosting for feature selection and logistic regression to identify key risk factors. The model showed good predictive accuracy, validated by receiver operating characteristic curves, calibration curves, and decision curve analysis. Key predictors included age, gender, prior surgeries, and nutritional status. This predictive tool aims to enhance clinical decision-making, reduce reoperation rates, and improve patient outcomes in colorectal cancer care.

INTRODUCTION

According to the World Health Organization, colorectal cancer is one of the most common malignant tumors of the digestive tract[1]. In 2018, there were more than 1.8 million cases of colorectal cancer globally, with a total of 881000 deaths — an average of 1 death out of every 10 cases[2]. Colorectal cancer is one of the top three cancer contributors to morbidity and mortality rates in the world[3]. Colorectal cancer poses a significant threat to the physical and mental health of the Chinese population. Early diagnosis of colorectal cancer in China is generally poor, and the majority of patients are in the middle-to-late stage of disease at the time of diagnosis[4]. Postoperative recurrence and metastasis of colorectal cancer are influenced by multiple factors such as lymph node metastasis, tumor type, growth location, and degree of infiltration. These factors are also key in determining the prognosis of patients with colorectal cancer[5].

Colorectal cancer is a serious malignant tumor and its treatment can include surgery, radiation therapy, chemotherapy, molecular-targeted therapy, immunotherapy, endocrinotherapy, and traditional Chinese medicine[6]. Currently, a combination approach based on surgery is the preferred strategy for the treatment of colorectal cancer[7]. Common surgical methods include radical surgery. However, in recent years, laparoscopy has been widely adopted due to its rapid recovery time, minimal trauma, and significant short-term efficacy[8].

Postoperative reoperation, particularly the rate of unplanned reoperation within 30 d, is an important indicator of surgical quality and has been adopted by the United States Centers for Medicare and Medicaid Services in its Physician Quality Reporting
System[9]. Due to the high morbidity and mortality of colorectal cancer, patients undergoing surgery are at risk of later reoperation. The percentage of postoperative unplanned reoperation in patients with colorectal cancer ranges from 3% to 11%[10,11]. The causes of reoperation include complications such as anastomotic leakage, bowel obstruction, and postoperative bleeding. Understanding the causes of reoperation helps improve patient prognosis. Despite improvements in surgical techniques and perioperative management, postoperative unplanned reoperation is still closely associated with complications[12]. These complications not only affect the short-term prognosis of the patient but may also apply surgical stress on the immune system, affecting postoperative outcomes. Unplanned reoperation is an independent predictor of a patient's mortality within one year of surgery[13].

Machine learning has great potential for disease risk prediction and diagnosis. In colorectal cancer, machine learning models can accurately predict the risk of undesired postoperative return to surgery by comprehensively analyzing multidimensional data on surgical approaches, and a patient's clinical characteristics and comorbidities[14]. The ability of such techniques to learn and adapt to new data means that their predictive accuracy continues to improve over time and data accumulation, reducing unnecessary reoperations, optimizing patient prognosis, and improving quality of life.

The purpose of this study is to establish and validate a model of colorectal cancer postoperative unplanned reoperation. This model combines multidimensional data including patient clinical characteristics; surgical modalities, and comorbidities, to improve prediction accuracy. This model will help physicians to more accurately assess patient postoperative risk, optimize treatment decisions, and reduce unplanned reoperation. This will ultimately improve patient prognosis and quality of life while reducing the economic burden of colorectal cancer on the healthcare system.

MATERIALS AND METHODS

Sample collection
Clinical data of patients with colorectal cancer admitted to the First Hospital of Wenzhou Medical University and Wenzhou Central Hospital from March 2020 to March 2022 were retrospectively collected. This study was approved by the Clinical Research Ethics Committee of the First Affiliated Hospital of Wenzhou Medical University and the Medical Ethics Committee of Wenzhou Municipal Central Hospital, No. KY2024-R016.

Inclusion and exclusion criteria

Inclusion criteria: (1) preoperative pathological findings confirmed the diagnosis of colorectal cancer[15]; (2) laparoscopic radical resection of the primary lesion; and (3) combined with distant metastases only radical resection of the primary lesion.

Exclusion criteria: (1) open surgery and intermediate open surgery; (2) intraoperative exploration found extensive metastases that could not be resected and only palliative surgery was performed; (3) resection of multiple bowel segments of both primary tumors or total or subtotal colectomy; (4) combined distant metastases were performed with simultaneous resection of the lesions; and (5) missing or incomplete clinical data.

Sample screening

We collected a total of 2948 patient records treated at the First Affiliated Hospital of Wenzhou Medical University and Wenzhou Central Hospital. According to the inclusion criteria, a total of 2484 samples met the requirements, and we excluded 440 samples. A total of 2044 samples were included. The patients with unplanned reoperation presenting at 30 d were assigned to the experimental group (n = 60). Among the 60 patients, 34 patients had anastomotic leakage, 21 patients had bowel obstruction, and 5 patients had abdominal cavity infection. The remaining patients were placed into the control group (n = 1984). To validate our model, we divided the patients into a training group (n = 1429) and a validation group (n = 615) based on a ratio of 7:3. Figure 1 depicts a flow chart of the process.
Clinical data collection
From the electronic medical records, we collected the clinical data of all patients, including age, gender, body mass index, history of hypertension, history of diabetes, history of stroke, history of laparotomy, preoperative hypoproteinemia, tumor site, history of preoperative radiotherapy, history of preoperative chemotherapy, tumor-node-metastasis (TNM) stage, American Society of Anesthesiologists (ASA) classification, surgical time, intraoperative bleeding, and preoperative prognostic and preoperative prognostic nutritional index (PNI).

Machine learning models
To efficiently screen feature variables associated with colorectal cancer postoperative unplanned reoperation, we used three different machine learning methods: support vector machine (SVM)\textsuperscript{[16]} least absolute shrinkage and selection operator (LASSO) regression\textsuperscript{[17]}, and extreme gradient boosting (XGBoost)\textsuperscript{[18]}.

The SVM method effectively distinguishes between two classes of data points (i.e., patients with or without unplanned reoperation) by finding an optimal hyperplane in a high-dimensional space. SVM is particularly effective when dealing with large datasets because it can work with high-dimensional feature spaces and nonlinear classification problems.

LASSO regression is particularly useful for feature selection as it reduces the coefficients of unimportant features to zero. This method limits the complexity of the model by adding a regularization term to avoid overfitting, while still identifying the most relevant features.

XGBoost is an integrated learning method based on decision trees, which improves prediction accuracy by constructing multiple models and combining them. It is an effective feature selection method as it optimizes the performance of the model through a gradient-boosting framework.
**Model evaluation tools**

To fully evaluate our unplanned reoperation, we used the following key statistical tools. The receiver operating characteristic curve (ROC) was used to assess the model's ability to discriminate between two types of outcomes (e.g., occurrence and non-occurrence of unplanned reoperation). The more diagnostic the model is, the closer the area under the curve (AUC) is to 1. We also used calibration curves to test the accuracy of the model's predicted outcomes. Ideally, the calibration curve should be close to 45 degrees, showing a high degree of agreement between predicted and actual values. The Hosmer-Lemeshow test (H-L test) was used to assess the fit of the model. A high P-value implies a good agreement between model predictions and actual observations. Decision curve analysis (DCA) was used to assess the utility of the model in clinical decision-making, as it identifies the thresholds at which the use of the model best improves patient care.

**Measurement of results**

Measurement of results: (1) The differences in clinical data between the control and experimental groups were compared; (2) SVM, LASSO, and XGBoost were used to screen for unplanned reoperation feature variables, and a Venn diagram was used to identify common feature variables; (3) Independent risk factors for postoperative unplanned reoperation were screened using logistic regression; (4) A nomogram was created based on the multifactorial logistic regression; (5) ROC curve, calibration curve, H-L test, and DCA were used to evaluate the differentiation, calibration, and clinical utility of the nomogram; and (6) Based on the risk coefficients, the risk scores of patients in the training and the validation groups were calculated. The differences in the risk scores of the patients were compared, and the predictive effect of the model was verified using the ROC.

**Statistical analysis**

Statistical analysis was carried out using SPSS 26.0 software. For normally distributed continuous data, used mean ± SD. Comparisons between groups were made using t-
tests. The chi-square test was used for count data. We screened all variables using SVM, LASSO, and XGBoost, and the common variables were screened using a Venn diagram. Multiple logistic regression analysis of the common variables was used to identify the independent risk factors. Then, we constructed a nomogram prediction model based on the selected independent risk factors using R software and the rms package. We obtained the calibration curve using Bootstrap and calculated the C-index. We also plotted the independent risk factors using ROC and calculated the AUC to validate the performance of the nomogram prediction model.

RESULTS

Comparison of clinical data
Comparison of the clinical data of the two groups showed that the number of patients in the experimental group aged $\geq$ 60 years, male, with a history of hypertension, a history of laparotomy and hypoproteinemia, and surgical time $\geq$ 240 mins was significantly higher than that of patients in the control group. The PNI of patients in the experimental group was also significantly higher than that of patients in the control group ($P < 0.05$, Table 1). The remaining variables were not statistically different ($P > 0.05$).

Machine learning models screening unplanned reoperation feature variables
We screened the unplanned reoperation feature variables using XGBoost, SMV, and Lasso methods (Figure 2). We found that XGBoost identified a total of 13 feature variables (Figure 3A), SMV identified 16 feature variables (Figure 3B), and Lasso identified 11 feature variables (Figure 3C). Using a Venn diagram, we found that the 3 methods screened 10 common characteristic variables: PNI, history of laparotomy, hypoproteinemia, age, TNM staging, history of hypertension, surgical time, gender, history of stroke, and ASA classification.

Logistic regression screening for independent risk factors for unplanned reoperation
We analyzed the 10 identified signature variables using multifactor logistic regression. The 10 signature variables were first assigned values (Supplementary Table 1). The resulting analysis revealed that age, gender, history of hypertension, history of laparotomy, hypoproteinemia, and PNI were independent risk factors impacting the likelihood of unplanned reoperation (Table 2, \( P < 0.05 \)).

**Establishment of a nomogram**

A nomogram prediction model was created based on the 10 predictors (age, gender, history of hypertension, history of laparotomy, hypoproteinemia, and PNI). The final prediction model equation was:

\[
\text{Logit (P)} = -6.8730575 + \text{Age} \times 1.108872309 + \text{Gender} \times 0.737188569 + \text{History of hypertension} \times 0.619231168 + \text{History of laparotomy} \times 0.619231168 + \text{History of hypertension} \times 0.917723145 + \text{Hypoproteinemia} \times 0.983183577 + \text{PNI} \times 2.48620524.
\]

The total score was obtained by summing the scores of each variable and finding the corresponding value on the "Total Score Axis". The value of the "Total Score Axis" was compared with the probability prediction line at the bottom of the nomogram to find the risk of postoperative unplanned reoperation (Figure 4).

**Evaluation of nomogram**

The differentiation, calibration, and clinical utility of the model were evaluated by four methods: ROC, calibration curve, H-L test, and DCA. The ROC analysis revealed that the AUC of the nomogram was 0.842, with 80.59% specificity, 76.67% sensitivity, and 57.26% Youden index (Figure 5A). This indicates that the model has a good degree of discrimination and can correctly distinguish the ending event from the non-ending event. Calibration curve analysis found that the nomogram’s calibration curve had a slightly poorer overlap, but generally went in the same direction (Figure 5B). The H-L test value was 8.588 (\( P = 0.378 \)). The DCA curve indicated that the unplanned reoperation net benefit rate was higher than other, *i.e.*, the blue line corresponding to
the threshold probability was located to the upper right of the All line (red line), indicating that the model had some clinical utility (Figure 5C).

Validation of nomogram
We divided the data into a training group and a validation group. The risk scores were calculated separately for both groups and then validated using ROC, calibration curve, H-L test, and DCA. As before, we compared the baseline information of patients in the training group with those in the validation group. The results showed that there was no statistically significant difference between the baseline characteristics of patients in the training group and the validation group ($P > 0.05$, Table 3). We then calculated the risk scores of the two groups, and the results showed that the risk scores of the patients who underwent unplanned reoperation were higher than those of patients in the non-reoperation group, both in the training and validation group ($P < 0.001$, Figure 6). Finally, we found that the AUC of patients in the training group and the validation group were 0.798 and 0.846, respectively. This suggests that the model can correctly differentiate between the outcome and non-outcome events.

Clinical validation of predictive modeling
To validate our model, we randomized the clinical data of 1 patient with unplanned reoperation. This patient was aged ≥ 60 years, male, had no history of hypertension, no history of laparotomy, hypoproteinemia, and his PNI was ≥ 43.76. The probability of occurrence was calculated for this patient ($45 + 30 + 0 + 0 + 39 + 100 = 216$). The results showed that the probability of the patient having unplanned reoperation was about 73% (Figure 8).

DISCUSSION
Treatment of colorectal cancer through laparoscopy allows comprehensive observation; clear peeling and resection of the lesion, as well as procedures such as hemostasis and lymph node dissection$^{[19]}$. Laparoscopy has a low impact on the patient’s abdominal
cavity, reduces postoperative pain, and promotes recovery of gastrointestinal function\textsuperscript{[20]}. However, despite the improved precision and safety of laparoscopy, unplanned reoperation remains a challenge for colorectal cancer outcomes\textsuperscript{[21]}. Reoperation not only prolongs the hospital stay and increases the financial burden of the disease, but also it affects the subsequent treatment plan and significantly increases the perioperative morbidity and mortality rate\textsuperscript{[22]}. Therefore, investigation of the causes and risk factors of postoperative reoperation in colorectal cancer has important clinical applications in reducing the rate of reoperation.

The absence of a standardized definition for unplanned reoperation has resulted in notable variations in the reported endpoint indicators for postoperative colorectal cancer across different medical centers. In a study by Feo and colleagues, covering 92 hospitals in China, the average reoperation rate for colorectal cancer surgeries was 9.7\%\textsuperscript{[23]}. Unplanned reoperation’s discrepancies were primarily attributed to disparities in medical resources and treatment approaches, which influence the risk of postoperative unplanned reoperations across various levels and regions of healthcare institutions. In contrast, the incidence of unplanned reoperations following laparoscopic surgery for patients with colorectal cancer was 2.94\%. Our results generally align with the laparoscopic reoperation rate for bowel cancer (approximately 3.8\%) reported by Speicher et al\textsuperscript{[24]}. These observations reinforce the efficacy and safety of laparoscopic surgery as a preferred treatment option for colorectal surgical interventions.

Patients with colorectal cancer undergoing abdominal surgery have a higher incidence of unplanned postoperative reoperation compared to other general surgical procedures\textsuperscript{[25]} due to their susceptibility to incisional and abdominal infections, venous thromboembolism, and perioperative complications\textsuperscript{[26,27]}. In addition, the inherent necessity of reconstructing abdominal organs during colorectal surgery increases the likelihood of postoperative complications, thereby increasing the likelihood of subsequent reoperations.

This study aimed to develop a predictive nomogram model. To construct this predictive model, we first employed three advanced computational techniques: SVM\textsuperscript{[28]},
LASSO\textsuperscript{20}, and XGBoost\textsuperscript{30}. These methods are known for their efficacy in managing high-dimensional datasets and their ability to identify critical variables in such datasets\textsuperscript{18}. Specifically, SVM excels at handling a wide range of datasets, LASSO mitigates overfitting through a penalty-based approach, and XGBoost is particularly effective at dealing with nonlinear relationships between data points. This multifaceted methodological framework facilitates a robust assessment of the significance of variables from multiple analytical perspectives\textsuperscript{31}. After identifying essential variables through these preliminary methods, we applied logistic regression analysis to investigate these identified variables. This analysis allowed us to identify independent risk factors that significantly impacted the probability of unplanned reoperation. Our findings suggest that age, gender, prior hypertension, history of laparotomy, hypoproteinemia, and PNI are key independent risk factors. These insights provide an understanding of the patient-specific risks associated with unplanned reoperation after colorectal cancer surgery and contribute to the clinical decision-making process.

Recent studies have identified the male gender as an independent risk factor for unplanned reoperation\textsuperscript{32,33}. This correlation is likely attributable to the male physiology, lifestyle habits, and adherence to postoperative rehabilitation protocols. Li \textit{et al}\textsuperscript{34} also highlighted age as a determinant, positing that elderly patients are at an elevated risk of undergoing unplanned reoperations, a conclusion that aligns with our observations. While not directly causing complications, the presence of comorbidities significantly influences surgical outcomes. Therefore, a comprehensive preoperative assessment and management of comorbid conditions are imperative to mitigating the likelihood of reoperation\textsuperscript{35}.

Numerous studies have substantiated the association between preoperative hypoproteinemia and the risk of unplanned reoperation. Saadat \textit{et al}\textsuperscript{36} recognized preoperative hypoalbuminemia as an independent risk factor in patients with rectal cancer, a finding corroborated by Michaels \textit{et al}\textsuperscript{37}, who linked malnutrition to increased risk of unplanned reoperation. Our study further confirms that patients with diminished preoperative albumin levels are at a heightened risk for such interventions.
The PNI is a crucial marker for evaluating a patient’s preoperative nutritional and immunological status. Lower PNI values often indicate suboptimal nutritional health, which can potentially compromise wound healing through impaired collagen synthesis and fibroblast proliferation\cite{38,39}. Improving patients' nutrition by enhancing albumin concentrations and optimizing PNI scores may significantly curtail the risk of unplanned reoperations following rectal cancer surgeries. Moreover, a history of prior abdominal surgeries is an independent risk factor for postoperative bowel obstruction following rectal resections\cite{40}. This suggests that such historical surgical interventions may lead to extensive abdominal adhesions, thereby complicating subsequent procedures and elevating the risk of complications.

Screening patients with high reoperation risk helps clinicians target perioperative observations and interventions, thus reducing unplanned reoperation and improving patient prognosis. In this study, we successfully predicted the incidence of unplanned reoperation through a constructed nomogram. The internal validation showed that the model was highly accurate and had good predictive efficacy.

However, there are some limitations to this study. First, the retrospective design of this study may lead to information and selection bias. Second, the lack of an external independent dataset for validation limits the generalizability and reproducibility of the model. Finally, the lack of long-term follow-up data in this study prevented assessment of the long-term outcomes of surgery and patient quality of life. In the future, we hope to use a prospective design to reduce bias, conduct external validation to enhance the generalizability of the model, and include long-term follow-up to assess the long-term impact of surgery. These improvements may more accurately predict colorectal cancer postoperative risk and improve patient outcomes and quality of life.

**CONCLUSION**

This study successfully established and validated a postoperative unplanned reoperation risk model for colorectal cancer. Through comprehensive analysis, we accurately identified independent risk factors affecting the risk of unplanned
reoperation: age; gender; history of hypertension; history of dissection; history of hypoproteinemia, and PNI. The application of the model in clinical practice can help to more accurately assess the postoperative risk of patients, thus optimizing treatment decisions, reducing the occurrence of unplanned reoperation, and improving patient prognosis and quality of life.

REFERENCES


<table>
<thead>
<tr>
<th>Rank</th>
<th>Source</th>
<th>Words</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.frontiersin.org">www.frontiersin.org</a></td>
<td>119</td>
<td>3%</td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.researchsquare.com">www.researchsquare.com</a></td>
<td>25</td>
<td>1%</td>
</tr>
<tr>
<td>3</td>
<td>doaj.org</td>
<td>21</td>
<td>1%</td>
</tr>
<tr>
<td>4</td>
<td>spandidos-publications.com</td>
<td>19</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>5</td>
<td>Daopeng Yang, Yan Wang, Bowen Zhuang, Ming Xu, Changxi Wang, Xiaoyan Xie, Gang Huang, Yanling Zheng, Xiaohua Xie. &quot;Nomogram based on high-frequency shear wave elastography (SWE) to evaluate chronic changes after kidney transplantation&quot;, European Radiology, 2022</td>
<td>16</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>6</td>
<td>f6publishing.blob.core.windows.net</td>
<td>15</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>7</td>
<td><a href="http://www.science.gov">www.science.gov</a></td>
<td>15</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>8</td>
<td>cnki.sris.com.tw</td>
<td>14</td>
<td>&lt;1%</td>
</tr>
<tr>
<td></td>
<td>Title</td>
<td>Authors</td>
<td>Conference/Journal</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>9</td>
<td>Application of Ensemble Learning Algorithm in Prediction and Control of Binder Quantity in Rotary Hearth Furnace</td>
<td>Chenyang Xia, Xiaowei Yang, Jun Yao, Xianxia Zhang</td>
<td>2023 5th International Conference on Industrial Artificial Intelligence (IAI), 2023</td>
</tr>
<tr>
<td>10</td>
<td>A proposed predictive model for advanced fibrosis in patients with chronic hepatitis B and its validation</td>
<td>Hiroki Nishikawa, Kunihiro Hasegawa, Akio Ishii, Ryo Takata et al.</td>
<td>Medicine, 2016</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>