Research Progress on the Application of Predictive Models in the Treatment of Colorectal Cancer

Chenyu Wu[1] and Tianqing Yan[2, *]

[1]School of Electronic Information Engineering Zhongshan College, Dalian Medical University, Dalian, 116000, China

[2]College of Sciences, Shanghai University, Shanghai, 200444, China

*Corresponding author: 22122108@shu.edu.cn

Abstract. Colorectal cancer (CRC), as a highly prevalent and lethal malignant tumor worldwide, has become a major public health issue requiring urgent global attention, posing a serious threat to the health of people across the globe. With the continuous advancement of medical research, predictive models for colorectal cancer treatment have been extensively studied and developed. This paper systematically reviews the construction methods, datasets used, and performance evaluation metrics of various mainstream colorectal cancer models both domestically and internationally. Through research and synthesis, it is found that while most models demonstrate good performance in internal validation, they also reveal numerous limitations in clinical practice. These limitations manifest as insufficient external validation, weak model generalization, and inadequate reporting standards and data sharing. Addressing these issues, this paper proposes corresponding solutions to provide practical reference for optimizing colorectal cancer treatment prediction models. This aims to advance the field toward greater scientific rigor and standardization.

Keywords: Colorectal cancer, Predictive model, Supervised learning, Unsupervised learning, Data-driven optimization.

1. Introduction

According to the 2020 global top-ten cancer statistics, colorectal cancer ranked second in mortality and third in incidence, and has become a major public-health problem that severely affects the health of people worldwide, imposes a heavy disease burden on society, and poses a substantial challenge to healthcare systems [1]. Among malignant tumors, CRC is characterized not only by high incidence and high mortality but also by a complex etiology influenced by genetic factors, lifestyle and the environment. With rapid advances in medical technology, predictive models related to CRC treatment - a current research hot spot - are being increasingly studied and implemented.

At present, both domestic and international research focuses on building CRC - related predictive models using machine-learning methods to improve predictive accuracy and clinical utility. However, the emphases differ: domestic studies tend to concentrate on predicting and reducing postoperative complications, the application of traditional Chinese medicine therapies, and the development of grassroots CRC screening tools; research abroad emphasizes integrating global datasets and deeply investigating the relationship between diet and CRC, while discussing how to strengthen interdisciplinary collaboration and resolve data-standardization issues to achieve earlier screening and treatment. These differences reflect varying priorities in medical resource allocation, research focus, and population characteristics across countries, and they provide a practical context for the diversified development of predictive models.

This review surveys six representative predictive models used in the CRC treatment process, summarizing their construction methods, datasets, evaluation criteria, and predictive performance. It provides a comprehensive analysis of how different models are applied at various CRC stages - including disease occurrence prediction, preoperative risk assessment, postoperative recurrence monitoring, and survival prognosis -with the aim of offering theoretical guidance for developing more optimized CRC treatment prediction tools. Such models can supply patients and clinicians with reliable preoperative and postoperative (recurrence and survival) predictions to enhance the scientific basis for clinical decision-making. For high-risk individuals, accurate prediction enables early detection and treatment, helps prevent delays that arise from underestimating risk, facilitates early intervention, and at the same time reduces patients' financial burden by decreasing the number of tests and conserving medical resources. Postoperative prediction can identify groups at high risk of recurrence, enabling timely monitoring and intervention and thereby reducing postoperative mortality. Finally, based on analysis the paper identifies limitations of current predictive models and discuss approaches to address key problems - such as insufficient data standardization and limited model generalizability - to contribute theoretical foundations for the further development and clinical application of CRC predictive models.

2. Overview of Mainstream Technologies

2.1. Supervised Learning

Supervised learning aims to learn a mapping between input and output variables from labeled training data. It assumes the existence of an unknown target function $f:X\rightarrow Y$, where $X\subseteq R^n$ is the input space and Y is the output space. Given a sample set.

$$D=\{(xi,yi)\}_{i=1}^{m} \tag{1}$$

The learning algorithm seeks to approximate f by producing a hypothesis $\hat{f}$ such that, for a new input x, the prediction $\hat{f}(x)$ is as close as possible to the true output y.

At the heart of supervised learning lies risk minimization. Structural risk minimization extends empirical risk minimization by adding a regularization term to suppress overfitting and thus improve the model's generalization ability.

Depending on the type of the output variable, supervised learning problems are usually divided into two categories. When the output is continuous, the task is regression (for example, linear regression). When the output is discrete, the task is classification (for example, logistic regression or random forest) [2–4].

Supervised models are commonly used in colorectal cancer research. For instance, a quadratic discriminant model was developed and optimized to create a "predictive model for populations with Traditional Chinese Medicine advantages in advanced colorectal cancer"; because it outperformed other discriminant methods in accuracy and stability, it was selected as the core prediction tool and deployed on a smart healthcare private cloud platform [5]. Other studies, using large-scale global dietary data, applied supervised methods such as neural networks, k-nearest neighbors, generalized linear models, and recursive partitioning to predict colorectal cancer [6].

In research predicting the risk of Low Anterior Resection Syndrome (LARS) after low rectal cancer surgery, multivariable logistic regression has been used to screen and validate independent risk factors. Building on univariate analysis and an initial LASSO screening, candidate variables were entered into a multivariable logistic regression. The results identified BMI, tumor diameter, anastomotic leakage, neoadjuvant therapy, and the distance from the tumor to the anal verge as independent risk factors for LARS ($P < 0.05$). A nomogram based on these variables was then constructed; evaluation on training and validation sets demonstrated good discrimination and predictive performance (AUC ≈ 0.846, accuracy 75.28%, precision 81.25%). This model supports individualized postoperative risk prediction and has clinical value for preoperative risk assessment and planning of interventions [7].

For survival prediction in colorectal cancer, models that capture time-dependent effects—such as Cox-Time and N-MTLR—have been applied to overcome the proportional hazards assumption of the traditional Cox model. Studies show that these approaches perform comparably to the Cox model in terms of discrimination and calibration, and they highlight factors such as R0 resection, TNM stage, and the number of positive lymph nodes as key predictors—underscoring their utility for clinical risk assessment and personalized treatment planning [8].

2.2. Unsupervised Learning

Unsupervised learning is another branch of machine learning characterized by the absence of explicit labels or target variables in the training data; instead, models learn by uncovering latent structure, patterns, or distributional regularities in the data. Common tasks include clustering, dimensionality reduction, and association rule mining, and unsupervised methods are widely used for data exploration, feature extraction, and anomaly detection.

For example, in machine-learning studies that predict colorectal cancer using global dietary data, t-SNE and UMAP were mainly employed to visualize high-dimensional data and reveal patterns, thereby demonstrating the separability of dietary features for CRC classification. PCA and factor analysis were used for dimensionality reduction and to probe latent structure, uncovering correlations among dietary nutrients and their links to CRC risk. The study also applied association-rule mining (e.g., Apriori) to discover frequent itemsets and potential rules connecting dietary features with colorectal cancer risk [6].

2.3. Data-driven Optimization

2.3.1 Feature Selection

Feature selection is a central step in machine learning and statistical modeling. Its goal is to identify, from the original set of predictors, a subset that carries the most information for the prediction task while removing redundant or noisy features. Proper feature selection can improve a model’s generalization, reduce computational cost, and increase interpretability.

For example, Zhao Lingyue et al. adopted a hybrid strategy combining manual (expert-driven) screening and importance-based selection. The former removes candidate variables based on medical knowledge and clinical plausibility, while the latter quantifies variable importance using methods such as gradient boosting machines and random forests to further discard ineffective or redundant features. The final set of 36 core features not only improved predictive performance but also enhanced the interpretability of the model results [9].

2.3.2 Statistical Analysis

Statistical analysis involves collecting, organizing, and examining data to reveal relationships, distributional properties, and latent patterns among variables. Several studies first use univariate analysis to identify candidate variables that show significant associations with the outcome of interest, providing a foundation for subsequent modeling [7].

2.3.3 Model Evaluation and Validation

Model evaluation and validation employ systematic procedures to assess a machine-learning model’s performance and robustness on both training and unseen data. Common approaches reported in the literature include 15-fold cross-validation resampling [6]. and bootstrap resampling for model assessment [10].

3. Commonly Used Datasets and Evaluation Criteria

3.1. Preoperative prediction

3.1.1 Small-sample datasets

Using the dataset from the Oncology Department of Xiyuan Hospital, China Academy of Chinese Medical Sciences as an example, 384 patients with stage IV colorectal cancer who received traditional Chinese medicine (TCM) treatment were identified through the hospital information management system and review of paper medical records. These patients were grouped into two categories: the TCM advantage group (n=209) and the non-advantage group (n=175).

The grouping method was based on data characteristics: gender, pathological classification, ECOG score, site of onset, site of metastasis, stage of western medical treatment at the time of initial diagnosis, TCM differentiation and classification, etc.

Secondary discriminant analysis was employed to evaluate the performance of four classification models (LDA, QDA, MDA, and FDA). The evaluation criteria included accuracy (proportion of cases correctly predicted by the model), sensitivity (true positive rate, proportion of cases correctly identified), specificity (true negative rate, proportion of cases correctly excluded), ROC (area under the curve) curves (relationships between true positives/sensitivity and false positives at different thresholds), and AUC (area under the ROC curve, ranging from 0.5 [no predictive power] to 1 [perfect prediction]). Ultimately, the QDA model demonstrated the best performance with an accuracy rate of 90.79%, test set sensitivity of 97.56%, and specificity of 82.86%. The training set AUC = 0.8286 and test set AUC = 0.9021 indicated good generalization performance [5].

3.1.2 Large-sample datasets

The dataset was compiled from 25 countries using public databases including the U.S. CDC (National Health and Nutrition Examination Survey) and the Global Dietary Database, comprising 3,520,586 records. After data processing, 109,342 individuals were retained for analysis, including 7,326 colorectal cancer patients and 102,016 non-colorectal cancer patients.

The evaluation metrics included accuracy, sensitivity analysis, specificity, Kappa coefficient (measuring the consistency between predicted outcomes and actual labels in classification models), and confusion matrix (assessing overall model consistency). Various supervised learning models (ANN, NN, KNN, GLM, recursive partitioning) were compared using 15-fold cross-validation with resampling methods, repeated ten times. All models demonstrated strong overall performance, with accuracy rates consistently exceeding 0.90. The artificial neural network (ANN) performed optimally, achieving a false positive rate of 1% for colorectal cancer and 3% for non-colorectal cancer [6].

3.2. Recurrence after surgery and survival prediction

3.2.1 Post-operative risk prediction

The data sets were selected from patients who had undergone rectal cancer-related surgery, and the relevant patients were strictly screened to obtain the optimal data set.

By constructing a variety of algorithms and using Lasso regression and cross-validation method of multi-factor Logistic regression, the optimal model was obtained, and the sensitivity and accuracy of the model were optimal [7-9].

3.2.2 Predictions after the fact

Patients who underwent colorectal cancer-related surgery were selected and divided into recurrent group and non-recurrent group to obtain the optimal data set.

A postoperative recurrence risk prediction model was established using BMI, diabetes status, and adenoma count as predictors. The model demonstrated an AUC of 0.6666 (95% CI: 0.5673–0.7659) with 56.60% sensitivity and 72.13% specificity in the training set. In the validation set, it showed an AUC of 0.6221 (95% CI: 4617–0.7824) with 84.61% sensitivity and 43.48% specificity. Although the calibration curve revealed a lack of consistency between predicted and actual probabilities, the model still holds practical clinical significance [10].

4. Current Research Level

4.1. Preoperative prediction model

Current preoperative prediction models for colorectal cancer have achieved test-set validation criteria (AUC). The integrated diagnosis and treatment strategies and personalized diagnosis and treatment strategies of Chinese and western medicine have been proposed. Through the data of many countries, the relationship between dietary patterns and colorectal cancer occurrence has been revealed, which provides a new perspective for the prevention of colorectal cancer.

Taking the research conducted at Yan'an University Affiliated Hospital as an example, the study classified patients diagnosed with colorectal adenoma through colonoscopy between January 2017 and December 2023 into case and control groups based on diagnostic criteria for advanced adenomas. Through LASSO and logistic regression analyses, eight variables including serum calcium levels and fecal occult blood positivity were identified as correlates with colorectal cancer development. These findings provide significant clinical implications for advancing early colorectal cancer screening.

4.2. Postoperative prediction model

Current postoperative prediction models for colorectal cancer have achieved test-set validation criteria (AUC). Taking the model developed by Zhao Lingyue et al. as an example, researchers collected data from 1,000 patients between January 2019 and January 2023. By constructing models using 10 machine learning algorithms and conducting cross-validation with SMOTE processing, they identified the random forest model as the best-performing approach. This optimization improved postoperative management strategies while revealing two critical factors influencing postoperative bleeding: "patient compliance with medical recommendations" and "gastric polyps." These findings have driven postoperative education interventions [9].

5. Problems and Possible Solutions

5.1. Lack of generalization and clinical applicability

Most CRC clinical prediction models have undergone internal validation but lack external validation across populations and regions, resulting in insufficient generalizability. Variables such as lifestyle factors remain without standardized measurement approaches. Moreover, the absence of clinical decision-level comparisons [11] exacerbates learning curve challenges, as real-world patients differ from trial participants, potentially leading to more pronounced errors during model training [12].

To address these challenges, Relevant organizations should promote multicenter external validation to enhance model generalization. Led by international organizations, hospitals across Asia, Europe, and the Americas should collaborate to establish a shared data platform for CRC standards. Independent predictive models should be developed for different regions including Asia, Europe, and Africa, while also accounting for variations in racial demographics and dietary patterns.

The relevant prediction system should be integrated and optimized to set the optimal intervention threshold for CRC, such as prioritizing surgery for people with a survival rate of less than five years. At the same time, attention should be paid to covering high-risk groups of CRC, such as people with heavy oil and salt diet and low-income groups.

5.2. Methodological bias and lack of transparency

The TRIPOD-Cluster report highlights several persistent issues in CRC model research: inadequate consideration of center effects and correlations in multicenter clustered data; information leakage during feature selection and cross-validation coupled with insufficient sample sizes; lack of external validation, model updates, and recalibrations; as well as non-disclosure of research code and data, along with failure to document critical details in reports [13].

To address these challenges, mandatory code open-sourcing should be enforced. By integrating clinical imaging data and developing predictive models that combine CRC imaging features with clinical metrics, it can enhance both sample information density and processing efficiency. An online learning system should be deployed to automatically update predictive functions with new cases, while quarterly model adjustments will be made using fresh clinical data.

By adopting federated learning to establish a new CRC prediction model, it aims to achieve interoperability of CRC data across different hospitals and institutions and attempt to utilize the FHIR open standard to address data exchange challenges arising from heterogeneous platforms, systems, and datasets, thereby advancing the development of CRC prediction models.

6. Conclusion

Colorectal cancer, a globally prevalent and highly lethal malignancy, holds significant clinical importance for developing predictive models during treatment. This systematic review examines the construction methods, datasets, and evaluation criteria of six mainstream predictive models from domestic and international studies. While most models demonstrate strong internal validation performance, practical applications still face challenges including insufficient cross-regional external validation, limited generalizability, and inadequate reporting standards with data sharing gaps. To address these issues, the paper proposes solutions such as multicenter external validation, establishing shared data platforms, deploying online learning systems, and implementing federated learning to enhance model scientific rigor and practical utility. Moving forward, with continuous advancements in medical technology and deepening interdisciplinary collaboration, colorectal cancer treatment prediction models will achieve greater precision and efficiency. This will enable more reliable preoperative predictions, postoperative recurrence assessments, and survival prognosis evaluations, facilitating early detection, timely intervention, and prompt treatment. Such progress will significantly improve patients' quality of life while advancing the development of colorectal cancer predictive models.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] 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(3):209-249.

[2] Vapnik V N. The nature of statistical learning theory. Springer, 1995.

[3] Bishop C M. Pattern recognition and machine learning. Springer, 2006.

[4] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer, 2009.

[5] Liu J, Wu B, Wang X, Luo D, Qiao C, Yang Y. Exploration on building a smart colorectal cancer diagnosis and prognosis prediction model based on private cloud platform. Beijing Journal of Traditional Chinese Medicine, 2023, 42(4): 369-375.

[6] Abdul Rahman H, Ottom M A, Dinov I D. Machine learning-based colorectal cancer prediction using global dietary data. BMC Cancer, 2023.

[7] Ma F. Construction of a risk prediction model for low anterior resection syndrome after rectal cancer surgery based on machine learning algorithms. Chengde Medical College, 2025.

[8] Yang X, Qiu H, Wang L, Wang X. Predicting colorectal cancer survival using time-to-event machine learning. J Med Internet Res, 2023.

[9] Zhao L Y, Luo Y J, Cheng X L, Cheng C S, Chen H L, Yu H W, & Wang X G. Interpretable machine learning prediction model for delayed bleeding risk following endoscopic resection of colorectal polyps. Huazhong University of Science and Technology (Medical Edition), 2025, 15(3): 13-20.

[10] Zhang P. Construction and validation of a predictive model for the occurrence and recurrence risk of advanced colorectal adenomas. Yan’an University, 2024.

[11] Usher-Smith J A, Harshfield A, Saunders CL, et al. Performance of FIT-based and other risk prediction models for CRC. Clinical Medicine, 2023.

[12] Wang Y, Li X, Chen H, et al. Evaluating generalizability of oncology trial results to real-world patients. Nature Medicine, 2024.

[13] Collins G S, Moons K G M, Riley R D, et al. Transparent reporting of clustered data prediction models: TRIPOD-Cluster checklist. BMJ, 2023.