scientific reports

Check for updates

OPEN

Development and validation of survival prediction tools in early and late onset colorectal cancer patients

Wanling Li[1,2], Jinshan Liu[3], Yuntong Lan[2], Dongling Yu[3] & Bingqiang Zhang[1]✉

This study aims to develop online calculators using machine learning models to predict survival probabilities for early- and late-onset colorectal cancer (EOCRC and LOCRC) over a 1- to 8-year period. We extracted data on 117,965 CRC patients from the published database spanning 2010 to 2021, divided into training and internal testing datasets. The data of 200 CRC patients from Chongqing Hospital of Jiangsu Province Hospital was used as the external testing dataset. We conducted univariate and multivariate regression analyses on the training dataset to identify key survival factors and develop predictive machine learning models. The models were evaluated using internal and external testing datasets based on AUC, accuracy, precision, recall, and F1 score. Web-based calculators were subsequently developed to predict survival curves for EOCRC and LOCRC patients under different treatment strategies. In the multivariate Cox regression analysis, 16 and 18 variables were independently significant survival factors for EOCRC and LOCRC, respectively. In the EOCRC group, the machine learning models achieved AUC values of 0.880 and 0.804 in the internal and external testing cohorts. For the LOCRC group, the machine learning models exhibited AUC values of 0.857 and 0.823 in the internal and external testing cohorts. The online calculators, powered by trained machine learning models, are accessible at https://eocrc-surv.streamlit.app/ and https://locrc-surv.streamlit.app/. These tools estimate survival probabilities for EOCRC and LOCRC patients under various treatment strategies and display the corresponding survival curves post-treatment over the 1- to 8-year period. This study successfully developed online calculators using machine learning algorithms to predict 1- to 8-year survival probabilities for EOCRC and LOCRC patients under various treatment strategies.

Keywords Colorectal cancer, Machine learning, Online calculators, Survival

Colorectal cancer (CRC) is the third most common kind of malignancy and the second most prevalent reason for death due to cancer[1,2]. Risk factors contributing to the incidence of colorectal cancer include age, hereditary factors, environmental factors, and lifestyle influences[3], such as diets high in fat, smoking, and excessive alcohol consumption. While the 5-year survival rate of CRC has improved and exceeded 65%[4,5], the survival for advanced CRC is less than 10%[6]. Laparoscopic surgical resection remains the standard of care for CRC, but chemotherapy and radiation therapy are frequently needed for advanced stages[7]. A reference tool is needed to evaluate the survival benefits of surgical resection, chemotherapy, and radiation therapy, which would assist in making decisions about the most appropriate treatment options.

To improve the precision medicine of CRC, studies classified CRC into two distinct subsets: early-onset colorectal cancer (EOCRC) and late-onset colorectal cancer (LOCRC)[8–10]. The cut-off age for EOCRC is not strictly defined but is generally set as 50 years[9]. This classification is crucial because EOCRC and LOCRC differ significantly in their molecular, genetic, and histopathological characteristics[11]. EOCRC, increasing among younger patients since the 1990s, tends to be more aggressive and invasive than LOCRC. These differences not

[1]Department of Gastroenterology, University-Town Hospital of Chongqing Medical University, Chongqing 401331, China. [2]Department of Gastroenterology, Chongqing Hospital of Jiangsu Province Hospital, The People's Hospital of Qijiang District, Chongqing 401420, China. [3]Department of Gastrointestinal Surgery, Chongqing Hospital of Jiangsu Province Hospital, The People's Hospital of Qijiang District, Chongqing 401420, China. ✉email: zhbingqiang@hospital.cqmu.edu.cn

only influence survival rates but also necessitate personalized treatment approaches. Researchers can develop more targeted and effective interventions by studying these EOCRC and LOCRC separately.

To date, the tumor-node-metastasis (TNM) staging system remains the standard method for staging malignant tumors, playing a crucial role in predicting cancer prognosis, formulating treatment plans, and evaluating the effectiveness of treatments[12]. However, the TNM staging system includes only three variables and overlooks other influential factors such as age and carcinoembryonic antigen (CEA) levels. Current studies on survival prediction often use nomograms to estimate the survival of CRC patients[13]. However, the performance of the nomogram is usually limited. With advancements in artificial intelligence, machine learning has begun to offer more precise diagnostic and prognostic tools by extracting extensive details from data. Despite it is potential, the application of machine learning in predictive studies, specifically for CRC cancer patients categorized by age, remains underdeveloped. In response, we have created web-based applications that employ machine learning models to predict survival in patients with EOCRC and LOCRC under different treatment strategies. These applications are validated with external datasets to enhance their reliability and accuracy.

Methods

Dataset description

This study utilized data from two CRC datasets. The primary dataset was obtained from the SEER database, a program that collects information on cancer incidence and survival. SEER covers approximately 28% of the US population. The SEER database is publicly available and de-identified, and the use of this data was exempt from local ethics committee approval. In addition to the SEER data, patient's information was collected from an independent cohort at Chongqing Hospital of Jiangsu Province Hospital. This external and independent cohort was named the CHJP cohort. The Ethical Review Committee of Chongqing Hospital of Jiangsu Province Hospital approved this part of the study with the approval number 20240005, which was conducted in accordance with the guidelines of the Declaration of Helsinki. The demographic data of SEER and CHJP cohorts are presented in Table 1.

Data collection

The clinicopathological data of patients diagnosed with CRC were retrieved using SEER*Stat software. The inclusion criteria for CRC samples in the SEER database included the following: (1) The code for selecting CRC is the International Classification of Diseases for Oncology codes (3rd edition, ICD-O-3) as C18.0 (Cecum), C18.2–18.9 (Ascending colon, hepatic flexure of colon, transverse colon, splenic flexure of colon, descending colon, sigmoid colon, overlapping lesion of colon, and colon, NOS), C19.9 (Rectosigmoid Junction), and C20.9 (Rectum Non-specified). (2) The selected histology subtypes in patients included adenocarcinoma (AD) (8140/3, 8144/3, 8201/3,8210/3, 8211/3, 8213/3, 8220/3, 8221/3, 8255/3,8260/3, 8261/3, 8262/3, 8263/ 3, 8310/3, 8323/3), mucinous adenocarcinoma (MA) (8480/3, 8481/ 3), and signet ring cell carcinoma (SRCC) (8490/3). (3) CRC was the first and only primary malignancy. (4) Patients' survival time should be over 0 months. (5) The age of patients should be more than 18. (6) We excluded patients whose reporting source was limited to autopsy or death certificate only. (7) We excluded patients who lacked complete clinicopathological information.

Eligible patients from the SEER cohort were randomly divided into the training cohort and internal testing cohorts, with a split ratio of 7:3. The training cohort was used to explore the prognostic factors and to construct machine learning models, the internal testing cohort was used for validation of the machine learning models. Additionally, external validation data were analyzed from patients diagnosed with CRC at Chongqing Hospital of Jiangsu Province Hospital from January 2013 to December 2020. Clinical and survival information was gathered through the clinical information system and telephone follow-up. This external testing dataset was named the CHJP cohort. The following information was collected: (1) Demographic parameters, including age at diagnosis, gender, and marital status; (2) Pathological parameters, including histology type, tumor location, tumor grade, CEA levels, TNM stage, and tumor size; (3) Treatment parameters including surgery, preoperative or postoperative treatment, including radiotherapy and chemotherapy; (4) Prognosis-related parameters, including OS. The data extraction process of SEER and CHJP cohorts is illustrated in Fig. 1.

Variable description

This study organized, categorized, and preprocessed the downloaded clinical baseline data from SEER. Age was a numeric variable ranging from 19 to 90. Gender was a categorical variable with categories of male and female. Marital status was defined into three subgroups: married, separated (divorced, separated, widowed), and single. Race was a categorical variable, including white, black, or others. Median household income was defined into four numeric values: '1 (Less than $59,999)', '2 ($60,000-$74,999)', '3 ($75,000-$99,999)', and '4 ($100,000 and above)'. The area distribution was preprocessed into two categories: rural and urban. Tumor size was a numeric variable measured in centimeters. The tumor site included the lower colon, overlapping colon, rectum, transverse colon, and upper colon. The histological type included two subgroups: AD or MA/SRCC. CEA status was recorded as normal or abnormal. Grade recode included I, II, III, and IV and were then transformed into 1, 2, 3, and 4. Based on the TNM classification system, T (Tumor) included T1, T2, T3, and T4; N (Nodes) included N0, N1, and N2; and M (Metastasis) included M0 and M1. Combined metastasis status (CMS) was defined as yes (metastasis in liver, lung, bone, or brain) or no. The treatment delay (the days between diagnosis and treatment) group included two subgroups: short waiting (≤ 28 days) and long waiting (> 28 days). Surgery included no surgery, partial colectomy, and proctocolectomy. Radiation treatment was categorized as yes or no, and chemotherapy was categorized as no or yes.

Characterisitics	SEER *N* = 117,965	CHJP *N* = 200
Overall survival (OS):		
Alive	76,644 (65.0%)	149 (74.5%)
Dead	41,321 (35.0%)	51 (25.5%)
OS time	4.21 (3.21)	5.02 (2.20)
Age	64.5 (13.6)	64.5 (10.6)
Gender		
Female	55,696 (47.2%)	92 (46.0%)
Male	62,269 (52.8%)	108 (54.0%)
Race		
Black	13,202 (11.2%)	NA
Other	12,966 (11.0%)	
White	91,797 (77.8%)	
Marital		
Married	67,460 (57.2%)	188 (94.0%)
Separated	29,115 (24.7%)	12 (6.0%)
Single	21,390 (18.1%)	
Income		
1 (Less than $59,999)	21,749 (18.4%)	NA
2 ($60,000-$74,999)	32,548 (27.6%)	
3 ($75,000-$99,999)	44,238 (37.5%)	
4 ($100,000 and above)	19,430 (16.5%)	
Area		
Rural	15,035 (12.7%)	NA
Urban	102,930 (86.8%)	
Histology type		
Adenocarcinoma (AD)	108,802 (92.2%)	183 (91.5%)
MA_SRCC	9163 (7.8%)	17 (8.5%)
Site		
Lower colon	30,492 (25.8%)	37 (18.5%)
Rectum	31,122 (26.4%)	110 (55.0%)
Transverse colon	12,040 (10.2%)	9 (4.5%)
Upper colon	44,311 (37.6%)	44 (22.0%)
Grade		
I	9204 (7.8%)	16 (8.0%)
II	88,282 (74.8%)	126 (63.0%)
III	17,971 (15.2%)	58 (29.0%)
IV	2508 (2.2%)	0 (0%)
CEA		
Abnormal	50,638 (42.9%)	81 (40.5%)
Normal	67,327 (57.1%)	119 (59.5%)
T stage		
T1	12,152 (10.3%)	7 (3.5%)
T2	16,794 (14.2%)	25 (12.5%)
T3	67,335 (57.1%)	96 (48.0%)
T4	21,684 (18.4%)	72 (36.0%)
N stage		
N0	61,738 (52.3%)	120 (60.0%)
N1	36,062 (30.6%)	43 (21.5%)
N2	20,165 (17.1%)	37 (18.5%)
M stage		
M0	100,492 (85.2%)	188 (94.0%)
M1	17,473 (14.8%)	12 (6.0%)
Tumor size	4.78 (2.49)	4.53 (1.93)
CMS		
No	51,964 (87.1%)	NA
Continued		

Characterisitics	SEER	CHJP
	N = 117,965	*N* = 200
Yes	7705 (12.9%)	
Surgery at local site		
No surgery	4159 (3.5%)	2 (1.0%)
Partial colectomy	107,890 (91.5%)	198 (98.5%)
Proctocolectomy	5916 (5%)	0 (0%)
Chemotherapy		
No	58,737 (49.8%)	84 (42.0%)
Yes	59,228 (50.2%)	116 (58.0%)
Radiation		
No	99,384 (84.2%)	192 (96%)
Radiation	18,581 (15.8%)	8 (4%)
Treatment delay		
Long (> 28 days)	38,640 (32.8%)	13 (6.5%)
Short (< 28 days)	79,325 (67.2%)	187 (93.5%)

Table 1. Clinicopathological characteristics in CRC patients in SEER and CHJP cohorts. The 7th edition of TNM staging information. *CMS* combined metastasis status.

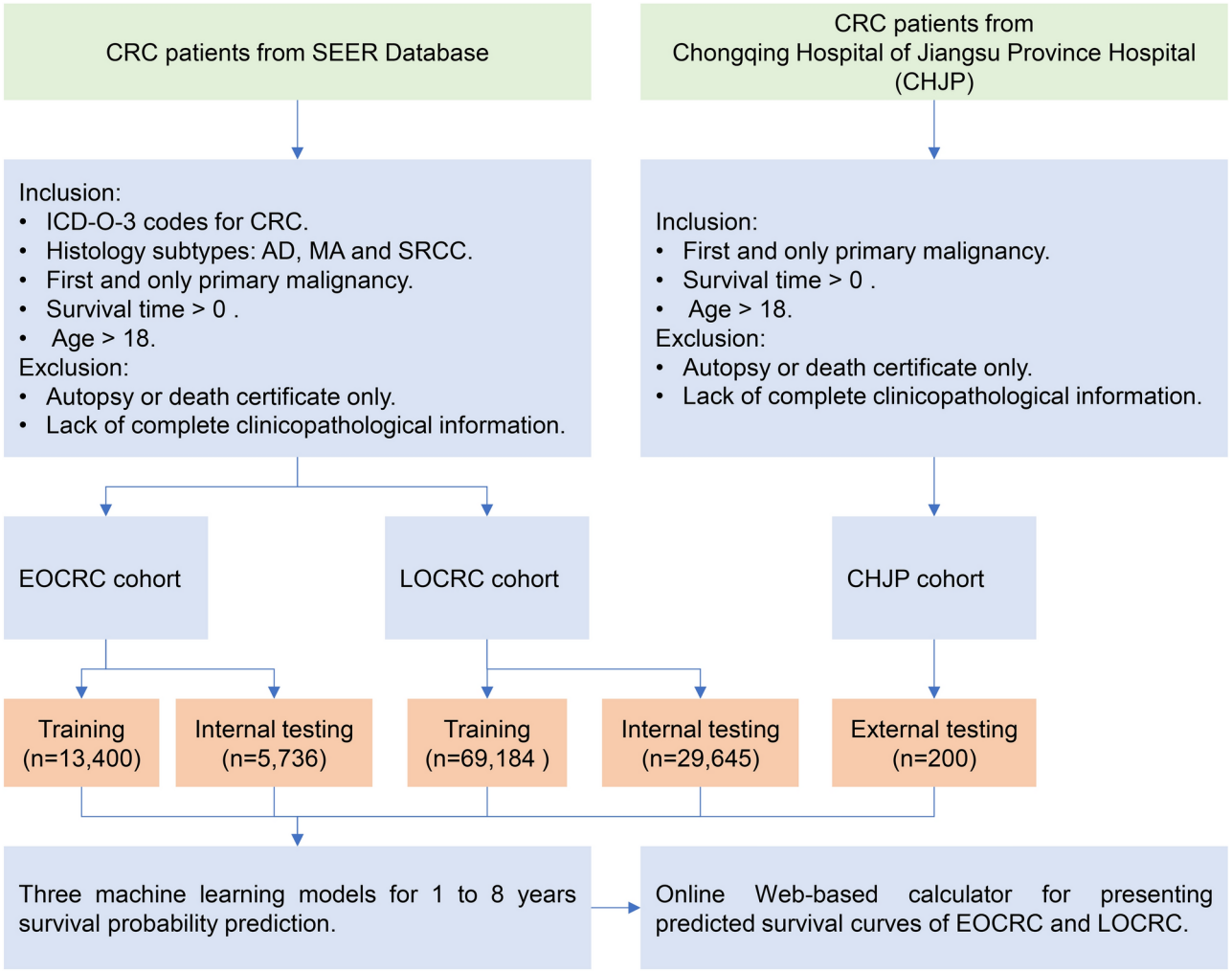

Fig. 1. The workflow and sample selection of this study.

Screening variables
Univariate analysis was firstly utilized to eliminate non-relevant variables and reduce overfitting in multifactor models. Variables that demonstrated statistical significance (p-value < 0.05) in the univariate analysis were considered in the multivariate analysis. Then, variables with p-value < 0.05 in the multivariate analysis were selected to develop machine learning models.

Model construction
This study aims to develop models that predict patient survival (alive or dead) at various time points (1, 2, 3, 4, 5, 6, 7, and 8 years). Patients were either alive or deceased based on survival status and duration. This classification led to potential imbalances at some time points. For example, over 90% of EOCRC patients were alive at the 1-year. This imbalance will limit the performance of machine learning models. For unbalanced cohorts, where the alive/dead or dead/alive ratio exceeded 5:1, we employed the SMOTE algorithm on the training set to adjust this imbalance. SMOTE, or Synthetic Minority Over-sampling Technique, is an advanced oversampling method used to address class imbalance by generating synthetic examples[14]. By creating synthetic samples, SMOTE helps to provide a more balanced class distribution and thus improves the performance of machine learning models on imbalanced datasets. It is noteworthy that SMOTE can only be used on the training set.

We utilized the training set to construct three machine learning models: random forest (RF), extreme gradient boosting (XGB), and gradient boosting (GB). Random forest is an ensemble method that reduces training variance by integrating multiple decision trees. XGB provides a high-performance implementation of gradient boosting, with optimizations for regularization and parallel tree construction. Gradient boosting sequentially constructs models, each improving on the errors of its predecessor, thus boosting overall predictive accuracy. To optimize these models, we applied five-fold cross-validation and rigorous parameter selection. Five-fold cross-validation involves randomly dividing the patient cohort into training and validation sets in a 4:1 ratio, and each segment serves once as the validation set. The model construction was conducted using the scikit-learn library from Python[15].

Model testing and model interpretability
The internal testing set from SEER and the external independent testing cohort from Chongqing Hospital of Jiangsu Province Hospital were utilized to evaluate three machine learning models using key performance indicators: area under the curve (AUC), accuracy, precision, recall, and F1 score. These metrics collectively provide a comprehensive assessment of a model's performance. In this study, we employed the permutation feature importance method to evaluate and identify the influence of variables within our machine learning model. By repeatedly shuffling each feature and assessing the impact on model accuracy, we calculated the average contribution of each feature to the model's predictive performance. We further identified the top six most influential features and calculated their relative importance compared to the most significant feature.

Survival analysis and web-based calculators
The models developed for the EOCRC and LOCRC cohorts were used to predict 5-year survival for each patient. Moreover, the predicted survival was used to estimate individual risk of death. Patients in the testing set were stratified into low-risk and high-risk groups based on median risk values, and survival curves were generated to illustrate the differences between these two groups. In addition, the models for EOCRC and LOCRC were used to construct web-based calculators. For these calculators, there are three surgery options (no surgery, partial colectomy, and proctocolectomy), two options for chemotherapy (yes or no), and two options for radiation therapy (yes or no). For EOCRC, radiation therapy was not a significant variable and was excluded from the model construction. Thus, six treatment combinations based on surgery and chemotherapy were available for EOCRC. For LOCRC, the inclusion of radiation therapy resulted in 12 treatment combinations based on the combinations of surgery, chemotherapy, and radiation therapy. The models generate survival rates for 1 to 8 years by combining the treatment options and clinical variables. Thus, our online calculator could produce survival curves for EOCRC patients with 6 treatment combinations and LOCRC patients with 12.

Results
Cohort information
This study obtained 117,965 CRC patients from the SEER database through screening (Fig. 1). The median age of the CRC patients was 65 years, with a mean survival/follow-up time of 4.2 years (Table 1). Patients were categorized into two groups based on an age cutoff of 50 years: EOCRC (≤ 50 years) and LOCRC (> 50 years). Each group was then randomly divided into training and internal testing cohorts. For the EOCRC group, the training and testing cohorts consisted of 13,400 and 5,736 samples, respectively. There were no significant differences (p-value > 0.05) in the selected variables between the training and internal testing cohorts, indicating comparable data sets (Table S1). Similarly, the LOCRC group included 69,184 samples in the training cohort and 29,645 in the internal testing cohort. The distribution of variables between the two cohorts is presented in Table S2.

Correlation of variable with clinical outcome
The univariate (Fig. 2A) and multivariate Cox (Fig. 2B) analyses of clinical variables in EOCRC identify significant correlations. Factors such as increased age, male gender, being separated or single, MA_SRCC, upper colon, higher tumor grades and stages, advanced TNM classification, and the presence of metastasis are associated with higher hazard ratios (worse survival). In contrast, certain racial groups, higher income, normal CEA levels, advanced surgical procedures, and undergoing chemotherapy are correlated with reduced hazards (better survival). Similarly, in LOCRC, univariate (Fig. 3A) and multivariate (Fig. 3B) Cox analyses reveal significant

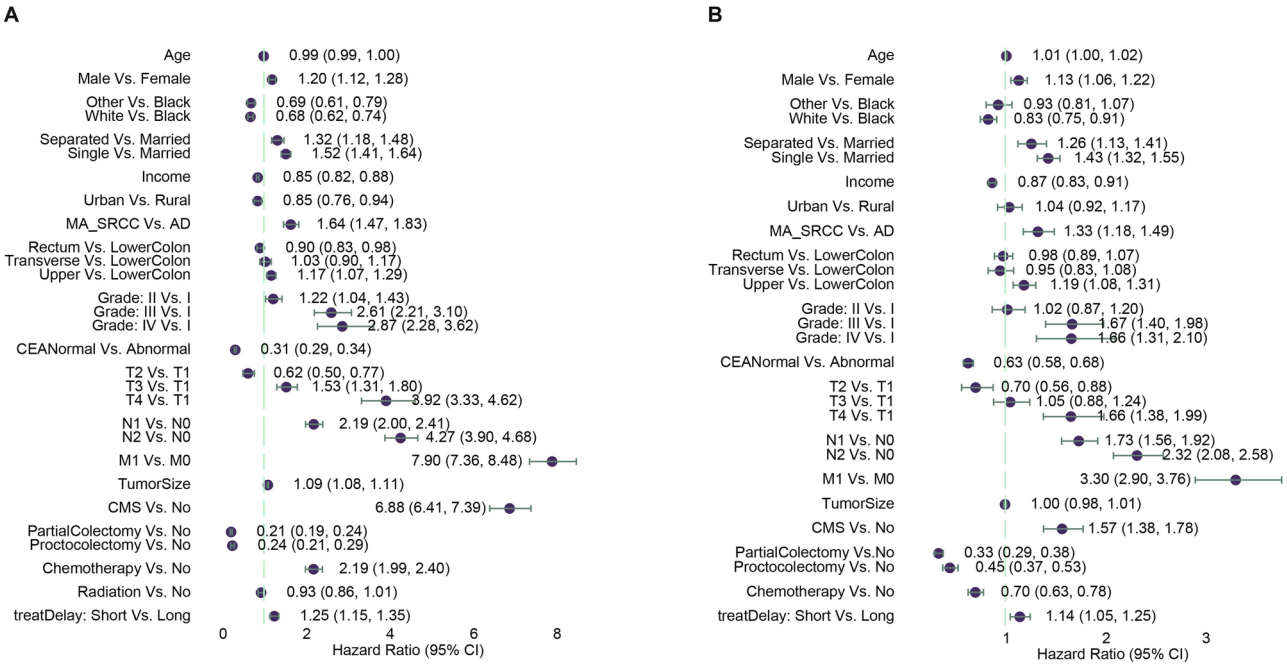

Fig. 2. Cox analyses of overall survival in EOCRC training set by forest plots. The forest plot shows univariate cox analyses of EOCRC (**A**) and multivariate cox analyses of EOCRC (**B**). The hazard ratio and its confidence interval were shown on these forest plots. If the confidence interval does not cross 1, the effect of the variable is considered significant. *CMS* combined metastasis status, *EOCRC* early-onset colorectal cancer.

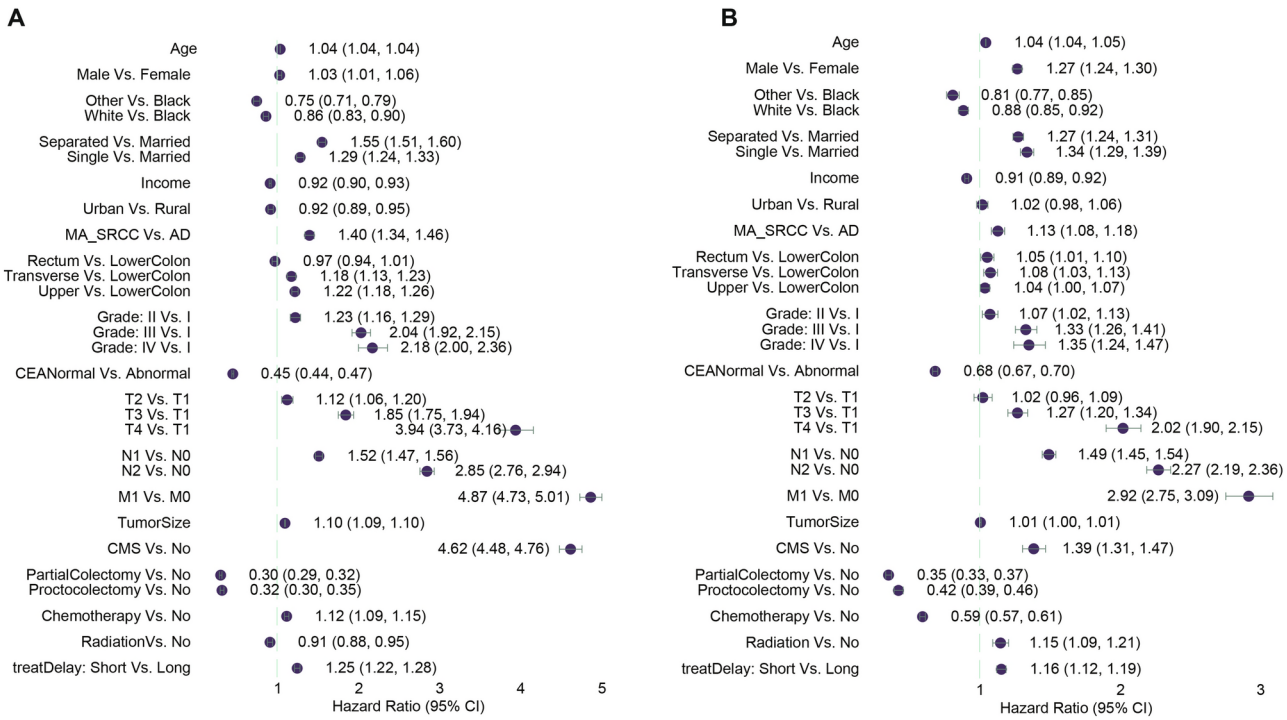

Fig. 3. Cox analyses of overall survival in LOCRC training sets by forest plots. The forest plot shows univariate cox analyses of LOCRC (**A**) and multivariate cox analyses of LOCRC (**B**). *CMS* combined metastasis status, *LOCRC* late-onset colorectal cancer.

associations. Factors leading to increased hazard ratios in LOCRC include increased age, male gender, being separated or single, MA_SRCC, higher tumor grades and stages, advanced TNM classification, larger tumor sizes, the presence of metastasis, use of radiation therapy, and longer waiting times before treatment. Conversely, belonging to higher income, normal CEA levels, advanced surgical interventions, and receiving chemotherapy are linked to decreased hazards. Then, variables with a p-value of less than 0.05 in the multivariate analysis were selected to develop machine learning models.

Establishing and evaluating ML models for estimating prognosis

We developed three machine learning models to predict EOCRC patients' survival from 1 to 8 years. The data for years 1 and 2 were notably imbalanced, with a living-to-deceased ratio exceeding 5:1 (Fig. 4A). Thus, we used the SMOTE technique to balance the training set at these years. We implemented five-fold cross-validation in the training set to determine the optimal hyperparameters and develop the most effective model. We generated predicted AUC, accuracy, F1 score, precision, and recall in evaluations using internal testing sets. The GB model demonstrated outstanding performance in predicting patient survival, achieving an average AUC of 0.880 (Fig. 4B), average accuracy of 0.842 (Fig. 4C), average F1 score of 0.707 (Fig. 4D), average precision of 0.744 (Fig. 4E), and average recall of 0.676 (Fig. 4F).

The distribution of the living-to-deceased patient ratio at different time points from LOCRC was provided (Fig. 5A). The GB model exhibited exceptional predictive performance regarding patient survival and achieved an average AUC of 0.857 (Fig. 5B), average accuracy of 0.803 (Fig. 5C), average F1 score of 0.750 (Fig. 5D), average precision of 0.780 (Fig. 5E), and average recall of 0.728 (Fig. 5F).

Evaluation of ML models in the external independent cohort from CHJP hospital

We collected clinical and prognostic information from 225 patients with CRC from Chongqing Hospital of Jiangsu Province Hospital to further validate the trained models. After excluding samples with unavailable or unknown information, 200 patients were selected as an independent testing cohort (Table 1). Figure 6A shows the number of living and deceased patients in the CHJP hospital. For XGB models trained on the SEER database for EOCRC, the validation results on the independent testing cohort showed that XGB showed the highest average AUC of 0.804 (Fig. 6B). The average values for the XGB model are as follows: an accuracy of 0.745 (Fig. 6C), an F1 score of 0.469 (Fig. 6D), a precision of 0.618 (Fig. 6E), and a recall of 0.399 (Fig. 6F).

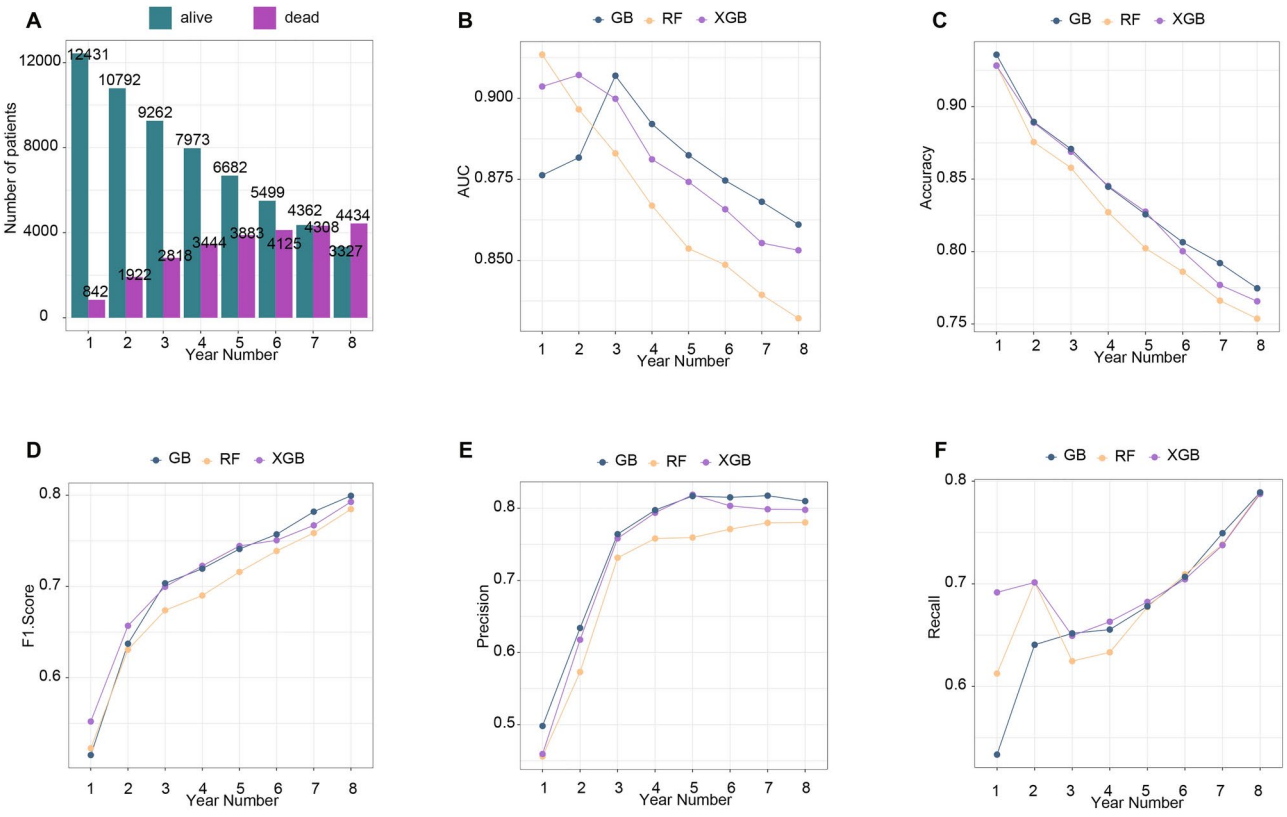

Fig. 4. Predicting the prognosis of EOCRC patients at various time points (1 to 8 years) in the internal testing cohort. (**A**) The number of alive/dead patients at different time points for EOCRC. Performance metrics of machine learning models evaluated in the internal testing cohort, including (**B**) AUC, (**C**) Accuracy, (**D**) F1 Score, (**E**) Precision, and (**F**) Recall. *RF* random forest (RF), *XGB* extreme gradient boosting, *GB* gradient boosting.

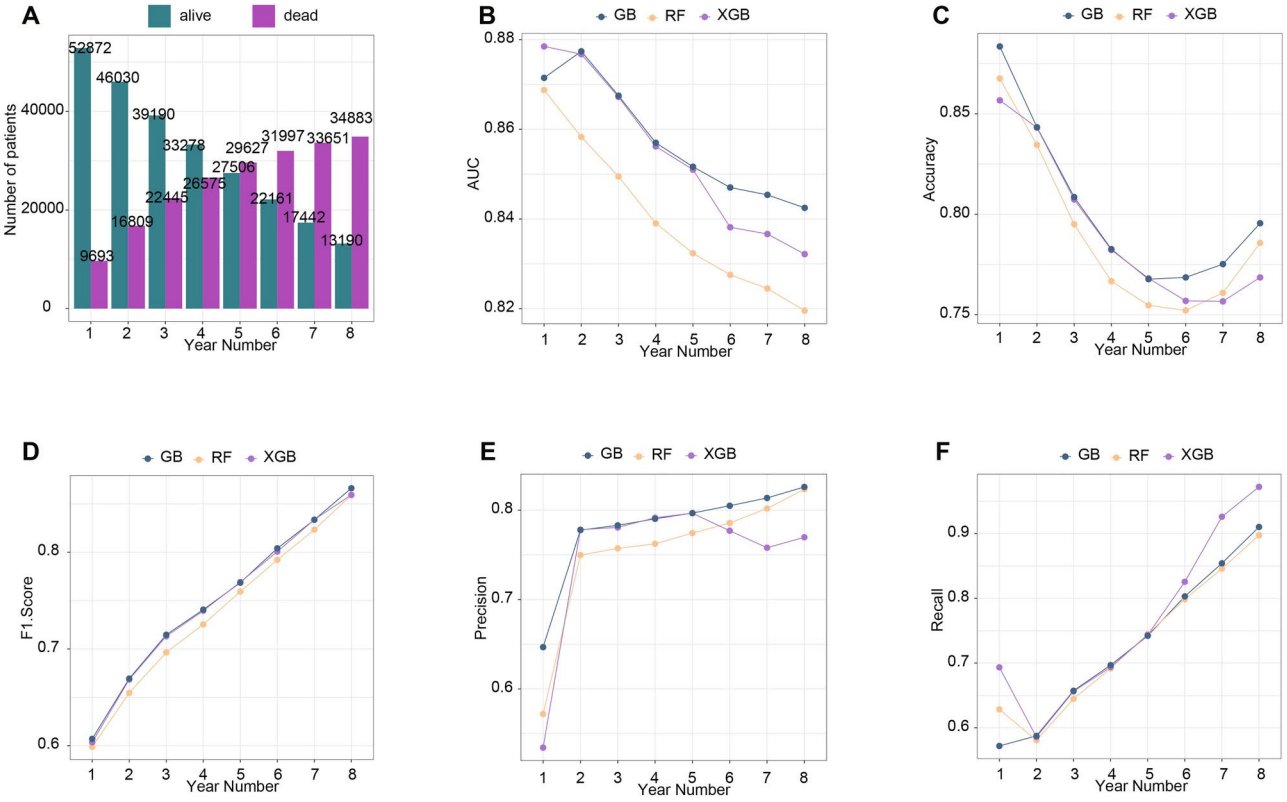

Fig. 5. Predicting the prognosis of LOCRC patients at various time points (1 to 8 years) in the internal testing cohort. (**A**) The number of alive/dead patients at different time points for LOCRC. Performance metrics of machine learning models evaluated in the internal testing cohort, including (**B**) AUC, (**C**) Accuracy, (**D**) F1 Score, (**E**) Precision, and (**F**) Recall.

Then, in the CHJP cohort, we evaluated the performance of models trained on the SEER database for LOCRC. The validation results on the independent testing cohort demonstrated that the GB model achieved the highest average AUC of 0.823 (Fig. 7A). Additionally, the GB model exhibited an average accuracy of 0.793 (Fig. 7B), an average F1 score of 0.612 (Fig. 7C), an average precision of 0.593 (Fig. 7D), and an average recall of 0.639 (Fig. 7E).

Relative importance of variables

The 5-year survival rate is a critical benchmark in cancer prognosis and thus was used for the variable importance evaluation. We analyzed the relative importance of six variables in three different models, GB (supFigure 1A), RF (supFigure 1B), and XGB (supFigure 1C), for EOCRC. Across all models, the TNM stage classification and surgery consistently ranked as the top predictors for EOCRC survival. Conversely, for LOCRC survival prediction, age was identified as the most significant variable in GB (supFigure 1D), RF (supFigure 1E), and XGB (supFigure 1F).

Survival analyses

Since the GB algorithm achieved the highest average in the independent testing cohort (Fig. 7A), it was selected for further analysis. We calculated the overall risk scores using the GB model for 5-year survival prediction in EOCRC patients. The overall risk scores were the predicted probability of death at 5 years. We categorized patients into low-risk and high-risk groups using the median score as the threshold. The Kaplan–Meier curves (Fig. 8A) demonstrate that individuals in the low-risk group had a significantly better prognosis than those in the high-risk group. Similarly, we calculated the overall risk scores using the GB model for 5-year survival prediction in LOCRC patients and plotted the corresponding survival curves (Fig. 8B). The results showed that the low-risk group had a significantly better prognosis than the high-risk group.

Web-based application development

We have developed user-friendly web applications to assist researchers and clinicians in utilizing our predictive models. These applications, shown in Fig. 8C-D, allow users to input the clinical characteristics of new patients and predict their survival probabilities and status based on data from patients with EOCRC or LOCRC. The two interactive manual interfaces, powered by trained models, estimate survival probabilities for EOCRC and LOCRC patients. They are accessible via the following links: https://eocrc-surv.streamlit.app/ and https://locrc-surv.streamlit.app/.

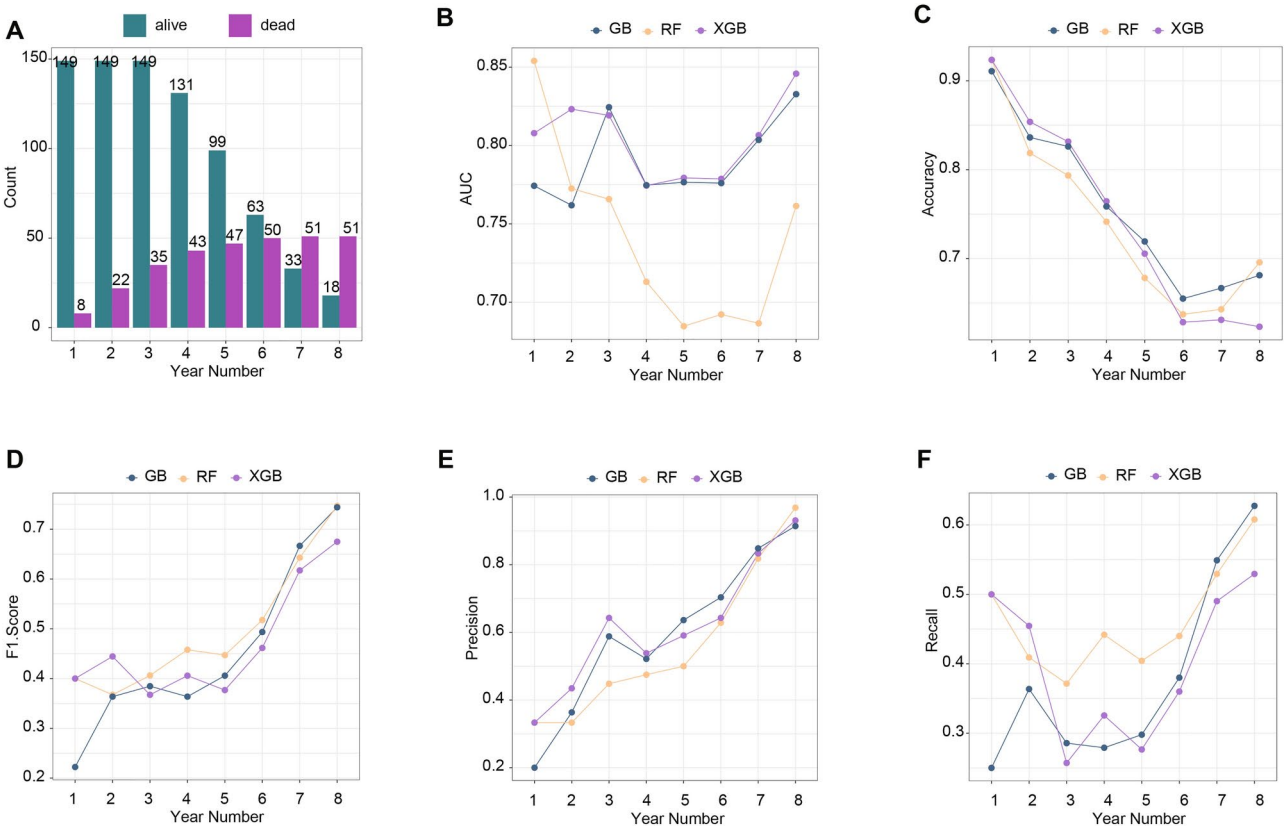

Fig. 6. Validation of models for EOCRC by external testing cohort. (**A**) Distribution of the alive/dead patient ratio at different time points. Performance metrics of machine learning models evaluated in the external testing cohort, including (**B**) AUC, (**C**) Accuracy, (**D**) F1 Score, (**E**) Precision, and (**F**) Recall.

Upon entering patient data, the calculators generate survival probabilities across various treatment options, ranging from 1 to 8 years. The output includes a survival curve plot and a data table displaying the predicted probabilities. These visualizations illustrate the time-dependent probability of survival, allowing users to compare survival curves for different treatment combinations on the same chart.

Discussion

This study constructed the machine learning algorithm-based prognostic models for EOCRC and LOCRC patients. The GB model outperformed other algorithms in the external testing cohort, demonstrating superior accuracy in predicting survival outcomes for both EOCRC and LOCRC. For EOCRC, TNM classification and surgery consistently ranked as the top predictors for EOCRC survival. Conversely, age was identified as the most significant variable for LOCRC survival prediction. The machine learning model-based prediction of survival curves under different treatment strategies showed promise for clinical use.

The incidence of EOCRC is increasing annually[16], while that of LOCRC is on the decline. EOCRC displays a distinct molecular profile that significantly differs from LOCRC, as evidenced by recent studies[17]. Typically diagnosed at later stages, EOCRC is characterized by more aggressive pathological features, delayed diagnosis, and unique genetic mutations[18]. Consequently, EOCRC and LOCRC should be regarded as distinct subsets and studied independently. Our analysis identified radiation and treatment delays as independent survival factors for LOCRC but not EOCRC. The AJCC Stage system and TNM classification emerged as the top predictors for EOCRC in the survival prediction models. Conversely, for LOCRC, age was identified as the most significant variable. Additionally, the interval between diagnosis and treatment, often defined as treatment delay, has been seldom explored in CRC studies. Our findings indicate that treatment delay is an independent factor affecting survival and should be considered when constructing predictive models.

Our study showcased distinct advantages of our model over recent studies in predicting the postoperative survival of colorectal cancer patients. For instance, a recent study developed a nomogram to predict survival for EOCRC, achieving 1-, 3-, and 5-year AUC values of 0.748, 0.733, and 0.720 in validation cohorts[19]. Another study constructed a nomogram model for colorectal cancer patients to predict 1-year OS undergoing neoadjuvant therapy, with AUC values of 0.765, 0.772, and 0.742 in different cohorts[20]. In comparison, our model for EOCRC and LOCRC demonstrated higher AUC values, reaching 0.880 and 0.857 in the testing cohorts. The model performance advantage highlights the superiority of machine learning models over traditional nomograms in this context.

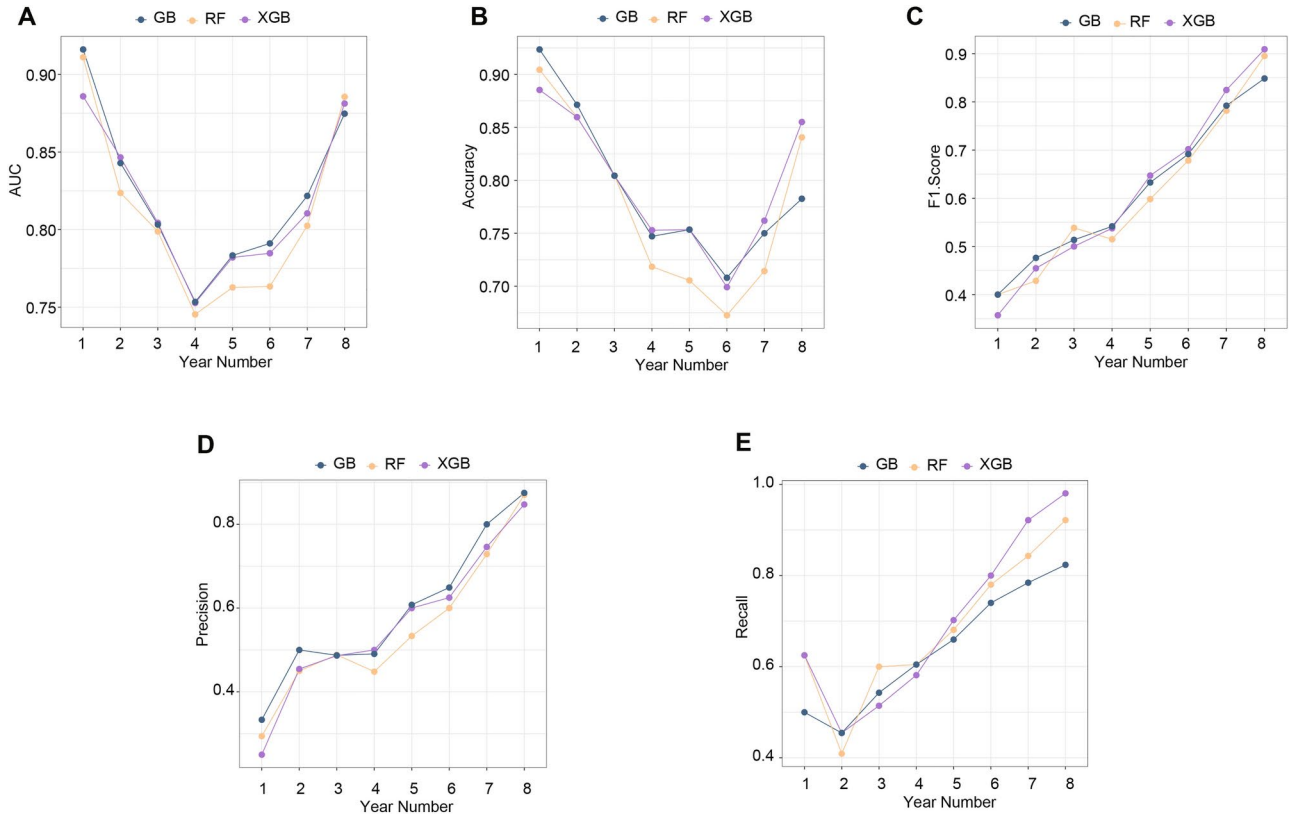

Fig. 7. Validation of models for LOCRC by external testing cohort. Performance metrics of machine learning models evaluated in the external testing cohort, including (**A**) AUC, (**B**) Accuracy, (**C**) F1 Score, (**D**) Precision, and (**E**) Recall.

The application of machine learning in colorectal cancer survival prediction is well-documented, with models achieving approximately 77% accuracy and AUC values nearing 0.86[21]. However, many studies do not provide user-friendly tools, such as applications or online calculators, making it difficult for patients and clinicians to apply these advanced models effectively. In contrast, our study not only offers precise predictive models but also includes online calculators that display survival probabilities for patients over a period of 1 to 8 years. This advantage enhances the clinical significance of our study, making it more practical for real-world use.

Machine learning models for predicting survival curves represent a significant advancement in predictive analytics, enabling a more comprehensive assessment of post-treatment outcomes based on initial clinical information and treatment-related variables. These models provide a personalized approach to patient care, allowing physicians to make more decisions. By utilizing our online survival probability calculators, available at https://eocrc-surv.streamlit.app/ for EOCRC and https://locrc-surv.streamlit.app/ for LOCRC, physicians can compare various treatment strategies and select the option that maximizes survival benefit for each patient.

Several limitations of this study need to be acknowledged. Firstly, while the models for EOCRC and LOCRC achieved AUC values of 0.880 and 0.857 in the internal testing dataset (American population), their accuracy may be overestimated. In the external testing dataset (Chinese population), the AUC values dropped to 0.804 and 0.823, respectively, indicating that the model's performance may vary across different populations. Secondly, our tool is not designed for direct treatment selection. For instance, there are no available data on rectal cancer patients for a precise evaluation of radiotherapy indications, which have evolved significantly in recent years. Furthermore, treatment decisions involve multiple complex factors beyond chemotherapy recommendations, including treatment regimens and sequencing, which cannot be fully identified and validated using an extensive population-based database. Finally, a comprehensive multi-center prospective study involving a large and diverse patient cohort is necessary to further assess the real-world applicability of our model. Therefore, while our tool is a valuable reference for survival prediction, its practical value in guiding treatment decisions should not be overestimated.

Conclusion

In this study, we developed accurate online calculators based on machine learning models to predict the survival of patients with EOCRC and LOCRC. The effectiveness of these tools was validated using independent testing cohorts. Future work will focus on enhancing model accuracy with larger datasets, integrating molecular biomarkers, and validating in multi-center trials for broader applicability.

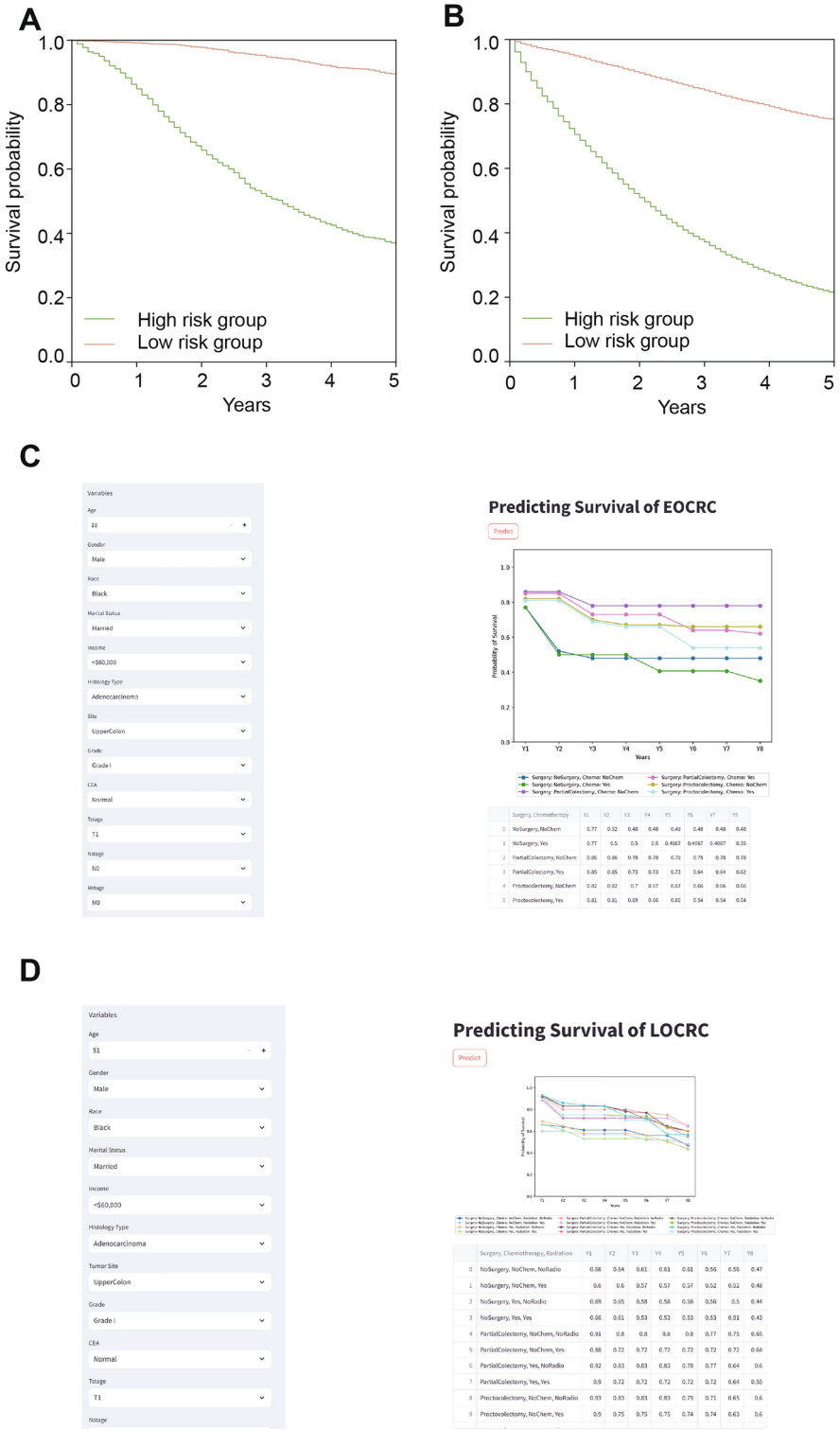

Fig. 8. Interactive Manual Interface for Predicting Survival Probabilities in EOCRC and LOCRC. (**A**) Kaplan–Meier curves showing survival outcomes for high- and low-risk groups in the internal testing cohort for EOCRC. (**B**) Kaplan–Meier curves for high- and low-risk groups in the internal testing cohort for LOCRC. (**C**) A demonstration of the online survival probability calculator for EOCRC. (**D**) A demonstration of the online survival probability calculator for LOCRC. Users simply need to input the relevant clinical variables and click the "Predict" button. The calculator will then generate survival probabilities across different treatment options, ranging from 1 to 8 years. The output includes a survival curve plot and a data table showing the predicted probabilities. To ensure the calculator is active, users may need to click the "Wake Up" button before use.

Data availability

Publicly available datasets were analyzed in this study which can be found here: https://seer.cancer.gov/. The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Received: 1 September 2024; Accepted: 20 March 2025
Published online: 15 April 2025

References

1. Ciardiello, F. et al. Clinical management of metastatic colorectal cancer in the era of precision medicine. *CA Cancer J. Clin.* **72**, 372–401. https://doi.org/10.3322/caac.21728 (2022).
2. Boatman, S., Nalluri, H. & Gaertner, W. B. Colon and rectal cancer management in low-resource settings. *Clin. Colon. Rectal. Surg.* **35**, 402–409. https://doi.org/10.1055/s-0042-1746189 (2022).
3. Rahiminejad, S. et al. Modular and mechanistic changes across stages of colorectal cancer. *BMC Cancer* **22**, 436. https://doi.org/10.1186/s12885-022-09479-3 (2022).
4. Brenner, H. & Chen, C. The colorectal cancer epidemic: challenges and opportunities for primary, secondary and tertiary prevention. *Br. J. Cancer.* **119**, 785–792. https://doi.org/10.1038/s41416-018-0264-x (2018).
5. Siegel, R. L. et al. Colorectal cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 145–164. https://doi.org/10.3322/caac.21601 (2020).
6. Maresca, C. et al. Smad7 sustains Stat3 expression and signaling in colon cancer cells. *Cancers* **14**, 4993. https://doi.org/10.3390/cancers14204993 (2022).
7. Kuipers, E. J. et al. Colorectal cancer. *Nat. Rev. Dis. Primers.* **1**, 15065. https://doi.org/10.1038/nrdp.2015.65 (2015).
8. Marx, O. M. et al. Transcriptome analyses identify deregulated MYC in early onset colorectal cancer. *Biomolecules* **12**, 1223. https://doi.org/10.3390/biom12091223 (2022).
9. Akimoto, N. et al. Rising incidence of early-onset colorectal cancer - a call to action. *Nat. Rev. Clin. Oncol.* **18**, 230–243. https://doi.org/10.1038/s41571-020-00445-1 (2021).
10. Gao, X. H. et al. Trends, clinicopathological features, surgical treatment patterns and prognoses of early-onset versus late-onset colorectal cancer: a retrospective cohort study on 34067 patients managed from 2000 to 2021 in a Chinese tertiary center. *Int. J. Surg.* **104**, 106780. https://doi.org/10.1016/j.ijsu.2022.106780 (2022).
11. Kirzin, S. et al. Sporadic early-onset colorectal cancer is a specific sub-type of cancer: a morphological, molecular and genetics study. *PLoS ONE* **9**, e103159. https://doi.org/10.1371/journal.pone.0103159 (2014).
12. Delattre, J.-F. et al. A comprehensive overview of tumour deposits in colorectal cancer: towards a next TNM classification. *Cancer Treat. Rev.* https://doi.org/10.1016/j.ctrv.2021.102325 (2022).
13. Zhang, Y. F., Ma, C. & Qian, X. P. Development and external validation of a novel nomogram for predicting cancer-specific survival in patients with ascending colon adenocarcinoma after surgery: a population-based study. *World J. Surg. Onc.* **20**, 126. https://doi.org/10.1186/s12957-022-02576-4 (2022).
14. Chawla, N. V. et al. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. https://doi.org/10.1613/jair.953 (2002).
15. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
16. Fiengo Tanaka, L. et al. The rising incidence of early-onset colorectal cancer. *Dtsch. Arztebl. Int.* **120**, 59–64. https://doi.org/10.3238/arztebl.m2022.0368 (2023).
17. Connell, L. C. et al. The rising incidence of younger patients with colorectal cancer: questions about screening, biology, and treatment. *Curr. Treat. Options Oncol.* **18**, 23. https://doi.org/10.1007/s11864-017-0463-3 (2017).
18. Garrett, C. et al. Early-onset colorectal cancer: why it should be high on our list of differentials. *ANZ J. Surg.* **92**, 1638–1643. https://doi.org/10.1111/ans.17698 (2022).
19. Yin, W. et al. Construction and validation of a nomogram for predicting overall survival of patients with stage III/IV early–onset colorectal cancer. *Front. Oncol.* **14**, 1332499. https://doi.org/10.3389/fonc.2024.1332499 (2024).
20. Han, Z. et al. The survival prediction of advanced colorectal cancer received neoadjuvant therapy—a study of SEER database. *World J. Surg. Onc.* **22**, 175. https://doi.org/10.1186/s12957-024-03458-7 (2024).
21. Buk Cardoso, L. et al. Machine learning for predicting survival of colorectal cancer patients. *Sci. Rep.* **13**, 8874. https://doi.org/10.1038/s41598-023-35649-9 (2023).

Author contributions

Wanling Li and Bingqiang Zhang: conceptualization, project administration, and funding acquisition. Wanling Li: software, formal analysis, data curation, visualization, and writing—original draft preparation. Jinshan Liu, Dongling Yu, and Yuntong Lan: resources. Wanling Li and Bingqiang Zhang: manuscript—reviewing and editing. All authors have read and agreed to the manuscript of submitted version.

Declarations

Competing interests

The authors declare no competing interests.

Ethics statement

For data from the SEER database, ethical review and approval were not required since the SEER database is publicly available and de-identified. For data from Chongqing Hospital of Jiangsu Province Hospital (The People's Hospital of Qijiang District), ethical approval was obtained from the Ethical Review Committee of Chongqing Hospital of Jiangsu Province Hospital (The People's Hospital of Qijiang District) with the approval number of 20240005 prior to commencing this study. The requirement for informed consent for retrospective study was waived by the Ethical Review Committee of Chongqing Hospital of Jiangsu Province Hospital (The People's Hospital of Qijiang District) because of the observational design and the anonymity of the patient's identity.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1

0.1038/s41598-025-95385-0.

Correspondence and requests for materials should be addressed to B.Z.