The Electronic Health Record-Based Artificial Intelligence Model for Predicting Long-Term Outcomes After Radical Surgery for Colorectal Cancer

Mariam Sh. Manukyan * , Valeriya I Pavlova , Maxim S. Kirsanov , Aydar Akhmetzyanov , Rukiyat Sh. Abdulaeva , Marianna O. Mandrina , Yana V. Belenkaya , Ivan S Stilidi , Tigran G. Gevorkyan , Sergey S. Gordeyev

Posted Date: 11 March 2026

doi: 10.20944/preprints202603.0700.v1

Keywords: colorectal cancer; cox regression; machine learning; CatBoost; survival analysis; AUC; TNM staging

 doi:10.20944/preprints202603.0700.v1

Article

The Electronic Health Record-Based Artificial Intelligence Model for predicting Long-Term Outcomes After Radical Surgery for Colorectal Cancer

Mariam Sh. Manukyan [1,*], Valeriya I. Pavlova [1,2,3], Maxim S. Kirsanov [4], Aydar Akhmetzyanov [4], Rukiyat Sh. Abdulaeva [1], Marianna O. Mandrina [1], Yana V. Belenkaya [1,5], Ivan S. Stilidi [1], Tigran G. Gevorkyan [1] and Sergey S. Gordeyev [1,3]

[1] N.N. Blokhin National Medical Research Center of Oncology of the Ministry of Health of Russia, 24 Kashirskoe Shosse, 115478 Moscow, Russia

[2] SBIH TO "MCMC "Medical City", Barnaulskaya Street, 32, 625041, Tyumen, Russia

[3] Tyumen State Medical University, Odesskaya Street, 54, 625023, Tyumen, Russia

[4] LLC “GlavGidroStroy”, 17-Y Mar'iny Roshchi Proezd, 4, Building 1, Room 18/19, 127521, Moscow, Russia

[5] Sechenov First Moscow State Medical University (Sechenov University), Trubetskaya Street, 8, Building 2, 119048, Moscow, Russia

* Correspondence: manukyanmariam6@gmail.com; Tel. +7-929-008-46-36

Simple Summary

Accurately estimating the risk of cancer recurrence after surgery is essential for personalizing follow-up care. However, traditional tools like the TNM staging system and logistic regression models have limited precision. In this study, we applied machine learning to develop a new predictive model using information already available in a patient's electronic health record. By analyzing data from over 7,000 patients, our model significantly outperformed traditional methods in forecasting both cancer recurrence and patient survival. The model utilizes 17 routinely available clinical variables, including blood test results and tumor characteristics, and can help clinicians identify patients at higher risk who may benefit from more intensive therapy and closer postoperative surveillance.

Abstract

Background: Accurate prediction of outcomes in colorectal cancer (CRC) is essential for personalized treatment. Conventional prognostic tools, including TNM staging, have limited accuracy. Machine learning (ML) may better capture complex prognostic patterns. Methods: In a retrospective multicenter cohort of 7,253 non-metastatic CRC patients after radical surgery, we compared prognostic accuracy for predicting recurrence and mortality using: a baseline TNM stage model; a logistic regression model with six clinicopathological variables; and ML algorithms (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost) with hyperparameter optimization (Optuna) and iterative feature selection. Binary outcomes (recurrence and all-cause mortality at 1 and 3 years) were used for ML training. Performance was assessed using area under the ROC curve (AUC). Results: The stage-only model showed poor discrimination (weighted AUC: 0.541 for mortality, 0.528 for recurrence). Logistic regression improved predictions (AUC: 0.759 and 0.645, respectively). Among ML models, CatBoost achieved the best performance. After iterative feature selection, the optimized CatBoost model utilizing 17 clinical variables demonstrated superior cross-validated AUCs of 0.81 for mortality and 0.84 for recurrence, consistently outperforming both baseline models across all time horizons. External validation on 1,452 held-out patients confirmed robustness with AUCs of 0.83 for mortality and 0.91 for recurrence. Conclusion: An optimized CatBoost model significantly outperforms traditional TNM staging and logistic regression in predicting recurrence and mortality in CRC using 17 routinely available variables. This

 doi:10.20944/preprints202603.0700.v1

parsimonious, data-driven tool offers improved individualized risk assessment for guiding post-operative management. Prospective validation is warranted.

Keywords: colorectal cancer; cox regression; machine learning; CatBoost; survival analysis; AUC; TNM staging

1. Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the second leading cause of cancer death worldwide [1]. Advances in screening and treatment have led to a high 5-year survival rate of 90-91% for localized disease [2]. Despite this progress, the risk of recurrence remains a major determinant of long-term outcomes, with modern estimates ranging from approximately 7% in stage I to 29% in stage III disease after curative surgery [3].

This persistent risk underscores the clinical challenge of identifying high-risk patients. Current surveillance protocols are largely uniform, potentially misallocating resources.[3] Accurate risk stratification is therefore essential to personalize follow-up. While traditional Cox regression provides a baseline, machine learning (ML) models may better capture complex prognostic patterns.

Personalized risk prediction may also guide adjuvant treatment decisions in the future. Recently, ctDNA detection has been recognized as a tool for selecting high-risk patients for adjuvant chemotherapy [4], and additional prognostic tools may further improve treatment personalization.

Previous publications have demonstrated the potential of machine learning-based prognostic models to accurately predict the risk of progression and survival in colorectal cancer patients after radical surgery. Studies using radiomics and clinical data have achieved C-indices up to 0.836 and accuracies up to 90.85% for recurrence prediction [5,6]. Models integrating clinical and genomic data have reported AUCs of 0.850–0.872 for survival prediction [7]. However, most of these data sources may have limited availability in daily practice. In contrast, data collected from electronic health records (EHR) are readily accessible and have demonstrated comparable efficacy [8,9]. However, outcomes across trials remain inconsistent due to the high heterogeneity of algorithms, clinical settings, and validation methods used. Furthermore, only a limited number of studies have directly compared machine learning models with traditional prognostic approaches, such as logistic regression or Cox regression, to demonstrate the superiority of ML-based algorithms [9].

We've conducted this study aiming to develop a ML-based prognostic model using only the readily available data from EHRs gathered in routine daily practice.

This study aimed to: develop a ML model to predict CRC recurrence and all-cause mortality in CRC patients undergoing radical surgery and to compare it with TNM staging and non-ML prognostic models.

2. Materials and Methods

2.1. Study Population and Data

This retrospective study analyzed data from EHRs of stage I-III CRC patients who underwent curative-intent surgical resection in 3 tertiary cancer centers during 2017-2024. The exclusion criteria were: distant metastases, multiple primary cancers. The depersonalized patient data was extracted from the hospital registries and united into a single dataset. A total of 143 data features were initially extracted for the database.

The study protocol was approved my the ethics committee of the N.N.Blokhin Russian Cancer Research Center (protocol number 01062024). The written informed consent requirement was waived considering the retrospective nature of the analysis.

 doi:10.20944/preprints202603.0700.v1

2.2. Predictor Variables and Target Outcomes

Predictors:

- Stage-Only Model: TNM disease stage.

- Logistic Regression Model: Univariate analysis was performed for all 52 extracted features, and the model was developed based on variables significantly ($p<0,05$) associated with the outcome. The final model retained six predictors: pT stage, pN stage, number of examined lymph nodes, perineural invasion, serum aminotransferase (AST) level, and international normalized ratio (INR).

- ML Model: Among 143 data features included in the dataset, 52 potentially relevant variables were selected, encompassing demographics, clinical characteristics, laboratory results, and treatment details. Further feature selection using iterative elimination was performed, resulting in a final model based on 17 variables. The full list of extracted clinicopathological variables and the final selected features are presented in Supplement Table S1.

- Target Outcomes: Two primary endpoints were analyzed:
  1. Recurrence: Defined as either locoregional or distant disease recurrence.
  2. Mortality: Defined as death from any cause.

Binary outcomes were constructed for 1-, 3-year prediction horizons for model evaluation.

2.3. Statistical and Machine Learning Analysis

2.3.1. Stage-Only & Cox Models:

Time-dependent AUC was calculated for each horizon (1, 3 years). The overall integrated discrimination was summarized using a weighted mean AUC, where weights were proportional to the number of events in each interval.

2.3.2. Machine Learning Model:

Missing Data Handling and Imputation Strategy

Missing values were addressed within the cross-validation loop to avoid information leakage.

For numeric and binary variables, we performed feature-wise selection of imputation strategies, evaluating several simple imputation rules (mean, median, most frequent, and a constant value) and a neighbor-based method (k-nearest neighbors imputation) as an alternative global strategy. KNN-based imputation estimates missing entries using the mean value from the nearest samples computed with a distance metric that supports missing values.

The choice of imputation approach was guided by cross-validated loss: we first selected the best simple strategy for each feature, and we additionally compared the resulting per-feature scheme against KNN imputation to choose the overall best imputation pipeline.

CatBoost was treated separately because it supports missing numerical values natively. Specifically, CatBoost can process NaNs by treating them as extreme values (minimum or maximum) and explicitly considering splits that separate missing from non-missing values; the behavior is controlled via the nan_mode setting.

Model Training and Evaluation

We evaluated a set of commonly used machine-learning algorithms for binary risk prediction, including logistic regression, random forest, and gradient-boosted decision tree models (CatBoost, XGBoost, and LightGBM). The objective was to identify the best-performing approach under a unified training and evaluation protocol, while ensuring robust generalization.

Data Prepocessing

For each candidate algorithm, hyperparameters were optimized using Optuna, an automatic hyperparameter optimization framework that implements an imperative “define-by-run” API and

 doi:10.20944/preprints202603.0700.v1

supports efficient search and pruning strategies. Hyperparameter tuning was performed within the cross-validation scheme, using cross-validated loss as the optimization objective, and the best configuration was then used for model comparison and subsequent analyses.

After comparing candidate algorithms, CatBoost was selected as the final model based on cross-validated performance. In the final stage, we performed an iterative feature selection procedure driven by model-based feature importance: at each iteration, feature importance was computed, the least informative features were removed, and the model was retrained. Importantly, at each iteration we re-ran the full training pipeline, including re-optimizing hyperparameters and re-selecting the missing-value handling strategy when applicable, and recorded the resulting loss. This process was repeated until removing additional features no longer improved performance or led to degradation. After iterative feature elimination and model retraining, a final set of predictors was retained in the CatBoost model. Feature importance was estimated using model-based importance scores, reflecting each variable's contribution to the predictive performance.

Evaluation and Validation

Model performance was assessed using the area under the receiver operating characteristic curve (AUC) for predicting recurrence and all-cause mortality at 1 and 3 years post-surgery.

Of the total 7,253 patients, 5,801 were used for machine learning analysis. No additional train-test split was performed; instead, five-fold cross-validation was applied on the entire cohort of 5,801 patients. Within each fold, hyperparameter tuning and iterative feature selection were performed on the training portion, and performance was evaluated on the held-out fold. The final reported AUC represents the average across all five folds.

The remaining 1,452 patients from the original database were not exposed to the model during cross-validation and were held out for external validation to assess generalizability.

3. Results

3.1. Patient Characteristics

The study cohort for comprehensive analysis comprised 7,253 patients. Among them, 3,639 (50.2%) were female. The distribution of disease stages showed a predominance of stage III (3,228 patients, 44.5%), followed by stage II (2,890 patients, 39.8%). At the time of final analysis, a recurrence event was documented in 693 patients (9.6%), and death from any cause was recorded for 1,344 patients (18.5%). Patient characteristics are summarized in Supplement Table S2.

3.2. Performance of the Stage-Only Model

Using a large cohort (n=7,253), a model based solely on TNM stage demonstrated very limited discriminatory ability for predicting mortality (AUC = 0.551) and recurrence (AUC = 0.552) over 3 years (Table 1). This established a performance baseline.

Table 1. Prognostic Performance of the TNM Stage-Only Model.

Endpoint	Time horizon	AUC
Mortality	1 year	0,54
	3 years	0,56
Recurrence	1 year	0,53
	3 years	0,55

3.3. Performance of the Logistic Regression Model

Logistic regression analysis was performed to develop a prognostic model for both recurrence and all-cause mortality. Initially, univariate analysis was conducted separately for recurrence and

 doi:10.20944/preprints202603.0700.v1

mortality at each time point (1 and 3 years). This revealed that different variables were significantly associated with each outcome and time horizon.

Multivariate analysis was then performed, but the resulting models showed inconsistent performance. For recurrence, three variables remained significant, and while the model performed well for overall recurrence, it failed to maintain predictive accuracy at individual 1-, 3-year time points. For mortality, only two variables were significant in multivariate analysis, and the resulting model showed poor predictive performance across all time points.

Given these limitations, an alternative approach was adopted. Rather than relying on multivariate selection, we performed iterative backward elimination, removing one variable at a time while monitoring model performance. The final model was selected based on the optimal balance between parsimony (fewest predictors) and predictive accuracy for both outcomes across all time points.

The final logistic regression model included six variables: pT stage, pN stage, number of examined lymph nodes, perineural invasion, AST, and INR. This single model was used to predict both recurrence and mortality, overall and at each time point (1 and 3 years).

On the entire cohort, the model demonstrated good predictive performance with an overall AUC of 0.7 for both recurrence and mortality. For recurrence, AUC values reached 0.85 and 0.87, at 1, 3 years, respectively. For mortality, AUC values were 0.82 and 0.85 at the same time points.

To further validate model performance, we subsequently split the dataset into a training cohort (70%) and an internal validation cohort (30%). The model was retrained on the training set and validated on the test set. The results were consistent with the initial findings and are presented in a heatmap (Figure 1), confirming the robustness of the final six-variable model across different data splits and time horizons.

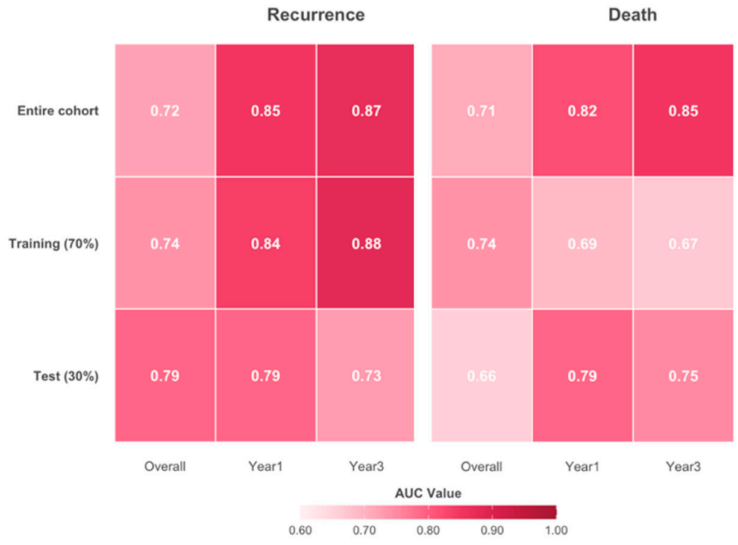

Figure 1. Logistic regression model performance (AUC) for recurrence and survival prediction.

3.4. Comparison and Performance of Machine Learning Models

Among the ML algorithms evaluated with five-fold cross-validation on the cohort of 5,801 patients, CatBoost consistently achieved the highest AUC for both endpoints (Table 2). It outperformed both the logistic regression model and other ML algorithms, with a cross-validated AUC of 0.770 for mortality and 0.749 for recurrence.

 doi:10.20944/preprints202603.0700.v1

Table 2. Cross-Validated ROC-AUC of Machine Learning Algorithms.

Algorithm	Recurrence AUC	Mortality AUC
CatBoost	**0.749**	**0.770**
Random Forest	0.738	0.748
XGBoost	0.736	0.747
LightGBM	0.725	0.741
Logistic Regression	0.692	0.713

3.5. Feature Optimization and Final Model Configuration

After evaluating multiple feature combinations, iterative feature selection was performed within the cross-validation framework. The optimal predictive performance was achieved with a refined set of 17 clinically relevant variables: absolute lymphocyte count, surgical approach, hemoglobin, weight, AST, leukocytes, number of positive lymph nodes, pretreatment CEA, platelets, height, number of examined lymph nodes, age at treatment initiation, total bilirubin, pT stage, NRAS mutation status, operation name, and disease stage.

The performance of the final optimized CatBoost model was assessed using two approaches: cross-validation on the entire cohort of 5,801 patients and external validation on the held-out cohort of 1,452 patients not exposed to the model during development.

When evaluated on the entire cohort using 43 features, the model achieved an AUC of 0.78 for recurrence and 0.81 for mortality. After feature selection, the 17-feature model demonstrated improved performance with cross-validated AUCs of 0.84 for recurrence and 0.81 for mortality. By time point, recurrence AUCs ranged from 0.82 to 0.85, and mortality AUCs ranged from 0.78 to 0.83 across 1, 3, and 5 years.

External validation of the 17-feature model showed even higher performance, with an overall AUC of 0.91 for recurrence and 0.83 for mortality. At individual time points, recurrence AUCs ranged from 0.84 to 0.91, while mortality AUCs ranged from 0.65 to 0.75.

The model achieved excellent discrimination for both primary endpoints, with cross-validated AUC values ranging from 0.78 to 0.85 across different prediction windows. Notably, the external validation demonstrated even higher performance for recurrence prediction, confirming the robustness and generalizability of the final 17-feature CatBoost model. The results were consistent with the initial findings and are presented in a heatmap (Figure 2)

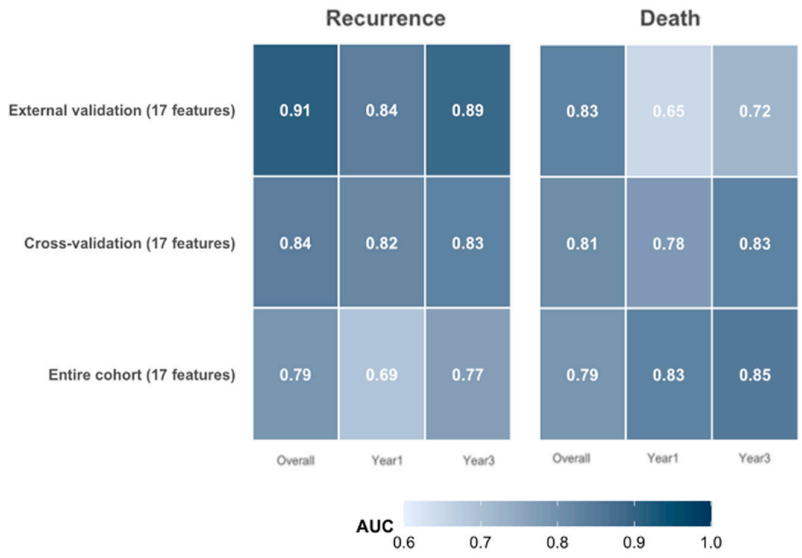

Figure 2. Machine learning model performance (AUC) for recurrence and survival prediction.

 doi:10.20944/preprints202603.0700.v1

3.6. Comparison of Logistic Regression and Machine Learning Models

Direct comparison of the final logistic regression model (6 variables) and the optimized CatBoost model (17 variables) revealed superior performance of the machine learning approach across all endpoints and time horizons.

The logistic regression model demonstrated good predictive ability on the entire cohort, with an overall AUC of 0.72 for both recurrence and mortality. At individual time points, it achieved AUCs of 0.85 and 0.87 for recurrence at 1, 3 years, respectively, and 0.82 and 0.85 for mortality. However, when evaluated on the internal validation set (30% hold-out), performance declined, with overall AUCs of 0.79 for recurrence and 0.66 for mortality, suggesting some degree of overfitting.

The CatBoost model, after iterative feature selection, achieved notably higher cross-validated performance with 17 features: overall AUCs of 0.84 for recurrence and 0.81 for mortality. By time point, recurrence AUCs ranged from 0.82 to 0.85, and mortality AUCs from 0.78 to 0.83. Most impressively, on external validation, the CatBoost model maintained or even improved its performance, achieving AUCs of 0.91 for recurrence and 0.83 for mortality overall, with recurrence AUCs reaching 0.84–0.91 across individual time points.

While the logistic regression model offers the advantage of interpretability with fewer variables, the CatBoost model consistently outperformed it across all metrics, particularly in external validation. The improvement was most pronounced for recurrence prediction, where the CatBoost model showed a 19% increase in overall AUC compared to logistic regression (0.91 vs. 0.72 on external validation). These findings underscore the value of machine learning approaches for capturing complex nonlinear relationships in clinical data, even when using a relatively modest set of routinely available variables.

4. Discussion

This study provides a direct, head-to-head comparison of three prognostic approaches for patients with colorectal cancer: a model based solely on TNM stage, a multivariable logistic regression model, and a machine learning model (CatBoost).

Our findings are consistent with the existing literature on machine learning-based prognostication in CRC. The strong performance of gradient boosting methods observed in our study is supported by multiple previous investigations. Kayikcioglu et al. [10] reported an AUC of 0.92 for recurrence prediction using CatBoost in a cohort of 396 patients that included laboratory parameters (CEA, albumin, platelet count, neutrophils, lymphocytes). Rodriguez et al. [11] achieved an AUC of 0.94 for recurrence prediction using XGBoost that incorporated longitudinal CEA measurements, demonstrating the added prognostic value of repeated biomarker assessments over time. Susič et al. [7], in a sample of 1,236 patients, reported the following AUC values for 5-year survival: logistic regression 0.87, XGBoost 0.86, LightGBM 0.86, and RF 0.85. Buk Cardoso et al. [12], in the largest cohort to date (31,916 patients), documented an AUC of 0.86 for XGBoost and 0.84 for RF. The consistency of these findings across diverse populations and healthcare settings underscores the robustness and generalizability of tree-based ensemble methods for survival prediction in CRC.

The advantage of gradient boosting algorithms becomes particularly evident when laboratory parameters are included in the model. In the study by Kayikcioglu et al. [10], the model incorporated multiple laboratory parameters: CEA, albumin, platelet count, neutrophils, and lymphocytes. These inherently continuous features often exhibit non-linear relationships with clinical outcomes and were effectively captured by the CatBoost algorithm. This supports the notion that gradient boosting methods are particularly well-suited for modeling complex interactions and threshold effects present in laboratory data. In our study, the inclusion of 17 heterogeneous features (clinical, demographic, and laboratory variables) similarly favored tree-based methods over linear approaches, with CatBoost achieving the highest AUC values (0.749 for recurrence and 0.770 for mortality). The superiority of machine learning algorithms over traditional regression models observed in our study aligns with the results reported by Tang et al. [13], where XGBoost (AUC 0.86) outperformed logistic

 doi:10.20944/preprints202603.0700.v1

regression (AUC 0.83), as well as with the work of Alinia et al. [9], where gradient boosting (AUC 0.96) and mboost (AUC 0.88) substantially surpassed Cox regression (AUC 0.54–0.71).

It is noteworthy that studies conducted on smaller cohorts, such as Alinia et al. [9] (N = 284, AUC 0.96 for recurrence and 0.88 for mortality) and Kayikcioglu et al. [10] (N = 396, AUC 0.92), reported higher absolute AUC values compared to our larger cohort. This discrepancy likely reflects the absence of overfitting in our models and the greater heterogeneity inherent to larger populations. In contrast, our results are comparable to studies by Buk Cardoso et al. [12] (N = 31,916, AUC 0.86).

In our analysis, XGBoost and RF demonstrated nearly identical performance (AUC 0.74 for recurrence and 0.747–0.748 for mortality), a pattern also observed in other large-scale studies [7,12]. This convergence may be attributed to the stabilizing effect of large sample sizes, which reduces variance between ensemble methods. However, in the study by Alinia et al. [9], conducted on the smallest sample (N = 284), RF exhibited a marked decline in performance (AUC 0.50 for both recurrence and mortality). This observation suggests that in small datasets, boosting may be more robust due to sequential error correction, whereas RF, by averaging multiple trees, may be more susceptible to noise and class imbalance. Logistic regression consistently underperformed relative to tree-based models (AUC 0.69–0.71), confirming the presence of non-linear dependencies and feature interactions in CRC data that cannot be adequately captured by linear models.

The final model incorporates 17 predictors, a number that falls within the optimal range identified in the literature, where most studies utilize between 10 and 20 parameters [8,19–21]. This balance is crucial for maintaining model parsimony without sacrificing predictive performance.

Comparison of our results with those of Susič et al. [7] revealed differences in the predictive importance of molecular biomarkers. In the study biomarkers such as KRAS, BRAF, and MSI did not rank among the most important prognostic features, being outperformed by clinical variables. In our study, conversely, NRAS mutation status was identified as a significant predictor. This discrepancy may be attributable to differences in cohort composition or feature engineering strategies. This finding underscores the need for cautious interpretation of biomarker importance and suggests the value of incorporating molecular data when available, although further validation on independent cohorts is required.

Notably, both aspartate aminotransferase AST and INR emerged as significant predictors in our logistic regression models. However, in the machine learning models, only AST retained its predictive significance across all algorithms, while INR did not maintain consistent importance. The persistent significance of AST across all modeling approaches suggests it captures a robust prognostic signal, potentially reflecting underlying hepatic dysfunction that could affect postoperative chemotherapy tolerability and dose intensity. However, due to the absence of detailed data on planned and administered adjuvant chemotherapy, this hypothesis requires further investigation. This finding highlights the potential value of incorporating liver function markers into prognostic models and warrants validation in future studies with comprehensive treatment data.

An important observation pertains to model behavior upon external validation. While the logistic regression model exhibited a decline in performance on the test set (AUC for recurrence decreasing from 0.74 to 0.69, and for mortality from 0.74 to 0.64), the CatBoost model not only maintained but improved its performance (recurrence AUC increasing from 0.84 in cross-validation to 0.91 in external validation). This finding indicates that the machine learning model possesses superior stability and generalizability—the capacity to maintain predictive accuracy when applied to novel, previously unseen data.

Furthermore, while external validation is infrequently performed in prediction model studies [19–21], we were able to validate our model on a substantial independent cohort (n = 1,452), where it sustained its high predictive accuracy.

An important direction for future improvement involves incorporating longitudinal biomarker data, as demonstrated by Rodriguez et al. [11], where dynamic changes in CEA over 6-month intervals achieved an AUC of 0.94, substantially exceeding models based on static measurements.

 doi:10.20944/preprints202603.0700.v1

uture studies should explore whether integrating repeated measurements of key biomarkers could enable dynamic risk stratification that updates over time.

In developing our model, we considered key aspects recommended by the updated PROBAST+AI tool [22], which is designed to assess the risk of bias and applicability of prediction models, including those based on artificial intelligence. Our study incorporated several of these criteria, including the use of multi-center data, clearly defined outcomes, external validation on an independent cohort, and predictor selection based on clinical relevance. A formal assessment of our model using the complete PROBAST+AI framework will be a crucial subsequent step to comprehensively evaluate its quality and potential for clinical implementation.

A key strength of this study is its utilization of routinely collected clinical data from electronic health records. Unlike approaches that rely on radiomics [15,16], pathomics [23], or genomics [24], our methodology does not require specialized equipment or additional costs, which may facilitate its integration into standard clinical workflows.

Looking ahead, the integration of our model into clinical practice could complement emerging molecular markers of minimal residual disease (MRD), particularly circulating tumor DNA (ctDNA). Detection of ctDNA has been demonstrated to identify patients at elevated risk of recurrence who may benefit from adjuvant chemotherapy [4]. While ctDNA provides dynamic, biology-driven risk assessment, our model offers static, baseline risk stratification using routinely available clinical variables. Future iterations of the model could incorporate ctDNA status alongside other multi-omics data, such as genomic or radiomic features, to create a more comprehensive prognostic instrument that captures both tumor biology and clinical phenotype. Such multi-modal approaches may further refine risk stratification and enable truly personalized treatment decisions.

This study has several limitations. Our primary focus was on model discrimination (AUC), and we did not conduct a detailed analysis of model calibration. The considerable heterogeneity in published studies also warrants caution when directly comparing absolute AUC values. Furthermore, as all patients were recruited from three regions within the Russian Federation, the model's performance should be validated in other geographic populations to confirm its generalizability.

5. Conclusions

In conclusion, our CatBoost-based model, developed with 17 routinely available clinical variables, demonstrates prognostic accuracy (AUC 0.84–0.91) comparable to the best-performing models in the literature and consistently outperforms traditional logistic regression. Its stability, strong performance on external validation, and reliance on readily accessible data make it a promising tool for integration into clinical practice to support personalized risk stratification for patients with colorectal cancer.

Author Contributions: Conceptualization, M.Sh.M., S.S.G. and I.S.S.; methodology, M.Sh.M., M.S.K. and A.A.; software, M.S.K. and A.A.; validation, M.Sh.M., V.I.P. and S.S.G.; formal analysis, M.S.K. and A.A.; investigation, M.Sh.M., R.Sh.A. and M.O.M.; resources, T.G.G. and I.S.S.; data curation, M.Sh.M., V.I.P. and Y.V.B.; writing—original draft preparation, M.Sh.M., M.S.K. and A.A.; writing—review and editing, V.I.P., R.Sh.A., M.O.M., Y.V.B., I.S.S., T.G.G. and S.S.G.; visualization, M.S.K. and A.A.; supervision, S.S.G. and T.G.G.; project administration, M.Sh.M. and V.I.P.; funding acquisition, S.S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been financially supported by The Analytical Center for the Government of the Russian Federation (Agreement No. 70-2024-000121 dd 29.03.2024. IGK 000000D730324P540002).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the N.N. Blokhin National Medical Research Center of Oncology (protocol code 01062024, date of approval 01 June 2024).

 doi:10.20944/preprints202603.0700.v1

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the analysis, as approved by the Ethics Committee.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors declare that Generative AI was used in the creation of this manuscript. During the writing process, ChatGPT 5.0 was used to optimization of English language expression and grammar checking. The author takes full responsibility for all content in this article. This statement is made to maintain academic transparency and integrity.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

AI	Artificial Intelligence
AST	Aspartate Aminotransferase
AUC	Area Under the Curve
CA19-9	Carbohydrate Antigen 19-9
CEA	Carcinoembryonic Antigen
CRC	Colorectal Cancer
ctDNA	Circulating Tumor DNA
EHR	Electronic Health Record
Hb	Hemoglobin
ICD	International Classification of Diseases
INR	International Normalized Ratio
LightGBM	Light Gradient Boosting Machine
LVI	Lymphovascular Invasion
ML	Machine Learning
MRD	Minimal Residual Disease
MSI	Microsatellite Instability
PNI	Perineural Invasion
PROBAST	Prediction Model Risk of Bias Assessment Tool
RF	Random Forest
ROC	Receiver Operating Characteristic
SD	Standard Deviation
TNM	Tumor/Node/Metastasis
TRG	Tumor Regression Grade
WBC	White Blood Cells
XGBoost	Extreme Gradient Boosting

 doi:10.20944/preprints202603.0700.v1

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2022;74(3):229-263.
2. American Cancer Society. Survival Rates for Colorectal Cancer. Cancer.org. Revised January 13, 2026. Accessed [Date].
3. Nors J, Iversen LH, Erichsen R, Gotschalck KA, Andersen CL. Incidence of Recurrence and Time to Recurrence in Stage I to III Colorectal Cancer: A Nationwide Danish Cohort Study. JAMA Oncol. 2023;9(12):1735-1744. doi:10.1001/jamaoncol.2023.5098
4. Masfarré L, Vidal J, Fernández-Rodríguez C, Montagut C. ctDNA to Guide Adjuvant Therapy in Localized Colorectal Cancer (CRC). Cancers (Basel). 2021;13(12):2869.
5. Zhu H, et al. Multi-center evaluation of machine learning-based radiomic model in predicting disease free survival and adjuvant chemotherapy benefit in stage II colorectal cancer patients. Cancer Imaging. 2023;23(1):74.
6. Volovat CC, et al. Machine Learning-Based Algorithms for Enhanced Prediction of Local Recurrence and Metastasis in Low Rectal Adenocarcinoma Using Imaging, Surgical, and Pathological Data. Diagnostics (Basel). 2024;14(6):625.
7. Susič D, et al. Artificial intelligence based personalized predictive survival among colorectal cancer patients. Comput Methods Programs Biomed. 2023;231:107435.
8. Kos FT, Cecen Kaynak S, Aktürk Esen S, Arslan H, Uncu D. Comparison of Different Machine Learning Models for Predicting Long-Term Overall Survival in Non-metastatic Colorectal Cancers. Cureus. 2024 Dec 14;16(12):e75713. doi: 10.7759/cureus.75713. PMID: 39811215; PMCID: PMC11730731.
9. Alinia S, et al. Predicting mortality and recurrence in colorectal cancer: Comparative assessment of predictive models. Heliyon. 2024;10(6):e27854.
10. Kayikcioglu E, et al. Machine learning for predicting colon cancer recurrence. Surg Oncol. 2024;54:102079.
11. Rodriguez PJ, et al. Using Machine Learning to Leverage Biomarker Change and Predict Colorectal Cancer Recurrence. JCO Clin Cancer Inform. 2023;7:e2300066.
12. Buk Cardoso L, et al. Machine learning for predicting survival of colorectal cancer patients. Sci Rep. 2023;13(1):8874.
13. Tang M, et al. Machine learning based prognostic model of Chinese medicine affecting the recurrence and metastasis of I-III stage colorectal cancer. Front Oncol. 2022;12:1044344.
14. Hu J, et al. Construction and validation of a progression prediction model for locally advanced rectal cancer patients. Front Oncol. 2023;13:1231508.
15. Ishizaki T, et al. Predictive modelling for high-risk stage II colon cancer using auto-artificial intelligence. Tech Coloproctol. 2023;27(3):183-188.
16. Nopour R. Development of Prediction Model for 5-year Survival of Colorectal Cancer. Cancer Inform. 2024;23:11769351241275889.
17. Li X, et al. Preoperative Albumin to Alkaline Phosphatase Ratio and Inflammatory Burden Index for Rectal Cancer Prognostic Nomogram. J Inflamm Res. 2024;17:11161-11174.
18. Zhang W, et al. Development and validation of an AI prediction model for lung metastasis in colorectal cancer. Eur J Surg Oncol. 2023;49(12):107107.
19. Jeon Y, et al. Machine learning based prediction of recurrence after curative resection for rectal cancer. PLoS One. 2023;18(12):e0290141.
20. Chen PC, et al. A Prediction Model for Tumor Recurrence in Stage II-III Colorectal Cancer Patients. Biomedicines. 2022;10(2):340.
21. Gupta P, et al. Prediction of Colon Cancer Stages and Survival Period with Machine Learning Approach. Cancers (Basel). 2019;11(12).
22. Moons KGM, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.

 doi:10.20944/preprints202603.0700.v1

23. Yu G, Sun K, Xu C, Shi XH, Wu C, Xie T, Meng RQ, Meng XH, Wang KS, Xiao HM, Deng HW. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nat Commun. 2021 Nov 2;12(1):6311. doi: 10.1038/s41467-021-26643-8. PMID: 34728629; PMCID: PMC8563931.
24. Radhakrishnan SK, Nath D, Russ D, Merodio LB, Lad P, Daisi FK and Acharjee A. Machine learning-based identification of proteomic markers in colorectal cancer using UK Biobank data. Front Oncol. 2025;14:1505675. doi: 10.3389/fonc.2024.1505675