# CEA-based machine learning methods for predicting recurrence and survival in colorectal cancer patients

Sukyong Yoon

Department of Medical Science

The Graduate School, Yonsei University

# CEA-based machine learning methods for predicting recurrence and survival in colorectal cancer patients

Directed by Professor Kyungsoo Park

The Doctoral Dissertation
submitted to the Department of Medical Science,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Sukyong Yoon

June 2024

This certifies that the Doctoral Dissertation of Sukyong Yoon is approved.

Thesis Supervisor: Kyungsoo Park

Thesis Committee Member#1: Joong Bae Ahn

Thesis Committee Member#2: Sang Joon Shin

Thesis Committee Member#3: Hyuk Hur

Thesis Committee Member#4: Dongwoo Chae

The Graduate School
Yonsei University

December 2023

# ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to Professor Kyungsoo Park, who not only helped me grow as a clinical pharmacologist throughout the years but also provided invaluable mentorship and emotional support as a senior in life. I also want to extend my sincere thanks to Professor Joong Bae Ahn, Professor Sang Joon Shin, Professor Hyuk Hur, and Professor Dongwoo Chae, who not only offered numerous pieces of advice but also contributed to various aspects of the research as the members of the thesis committee. And I am grateful to all the professors in the Department of Pharmacology who helped provide a desirable environment for research and education, enabling me to work in an excellent setting. I would like to acknowledge the faculty and staff at the Severance Clinical Trials Center for their support, as well as my colleagues in the Clinical Pharmacology Laboratory who were always there to lend a helping hand from the closest of distances. I will carry the lessons you've imparted throughout my life, and the bonds we've forged will remain with us as we step into the next chapter of our lives.

## TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

## ABSTRACT

### CEA-based machine learning methods for predicting recurrence and survival in colorectal cancer patients

Sukyong Yoon

*Department of Medical Science*
*The Graduate School, Yonsei University*

(Directed by Professor Kyungsoo Park)

**Objectives**: While colorectal cancer is the second leading cause of cancer-related deaths in developed countries and some patients still experience recurrence even after receiving the appropriate treatment, it is known that early diagnosis of recurrence improves the patient's prognosis. Nevertheless, currently there is no noninvasive approach available that enables early detection of recurrence. In this regard, this work was conducted to develop methods for early prediction of recurrence and survival in Korean colorectal cancer patients.

**Materials and Methods**: Our data consisted of 4,020 patients who underwent surgery and were diagnosed with stage I-III colorectal cancer at Severance Hospital (Seoul, Korea). From each patient, demographic information and clinical characteristics, including pre- and post-operative CEA levels, the number of infiltrated lymph nodes, the number of examined lymph nodes, tumor location, and age at the time of surgery, were collected as potential predictive variables for early detection of recurrence and survival in colorectal cancer. Additionally, another predictive variable named 'Slope’, which was derived from the blood levels of carcinoembryonic antigen (CEA), representing the slope of linear regression of CEA samples over the range from pre-recurrence up to approximately 1-year after surgery. Patients with a missing value for any of these variables were excluded. The analysis was conducted in two steps. In the first step, classification models were developed to predict recurrence status and survival status, respectively. In the second step, time-to-event models were developed to predict recurrence time and survival time, respectively. Then, given flexibility and scalability of machine learning, which does not require any specific form of a model and can be implemented based on the data available only, thus well suited for big data analysis such as retrospective studies based on electronic medical records, machine learning was used for model development. All data analysis and model building were performed using R software (ver 4.2.2) and its packages.

**Results**: Classification models were developed by testing various machine learning algorithms, including logistic regression, support vector machine, decision tree, random forest, gradient boost, XGboost, Light-GBM, and CatBoost. These models demonstrated Area Under the Receiver Operating Characteristic Curve (AUROC) values ranging from 0.87 to 0.92 for recurrence status and 0.87 to 0.89 for survival status. Among these models, the CatBoost model showed slightly better performance. Time-to-event models were developed using the random survival forest algorithm, resulting in AUROC values of 0.90 for recurrence time and 0.89 for survival time, respectively. In all developed models, the newly created 'Slope' variable was consistently selected as the most important predictor. For the time-to-event models developed, an R Shiny application was created to facilitate individual patient-level predictions.

**Conclusions**: This work demonstrated the feasibility of utilizing CEA in early detection of recurrence status, survival status, recurrence time, and survival time in colorectal cancer. The developed model achieved good predictive performance. It is hoped that the model and the R Shiny application developed will be helpful in assessing the prognosis of colorectal cancer patients in Korea.

Key words: colorectal cancer, machine learning-based models, recurrence, survival, carcinoembryonic antigen

# CEA-based machine learning methods for predicting recurrence and survival in colorectal cancer patients

Sukyong Yoon

*Department of Medical Science*

*The Graduate School, Yonsei University*

(Directed by Professor Kyungsoo Park)

## I. INTRODUCTION

Cancer ranks as the second leading cause of death in developed countries, with colorectal cancer being the second leading contributor, following lung cancer according to a publication in 2021[1]. In Global Cancer Statistics, it is reported that the worldwide number of colorectal cancer patients increased by approximately 2 million in 2020, resulting in a death toll of around 1 million[2]. Korean colorectal cancer patients show a similar trend, with a mortality rate of 17.4 per 100,000 people, ranking as the third highest, following lung cancer (36.4) and liver cancer (20.6)[3]. Unfortunately, some colorectal cancer patients experience recurrence even after receiving appropriate treatments such as resection and chemotherapy. It is well-known that early detection of recurrence through rigorous follow-up procedures can significantly enhance prognosis[4]. Hence, it is imperative to assess the recurrence risk in patients, identify high-risk groups for cancer recurrence, and facilitate intensive follow-up for these patients.

Over the years, numerous studies have been conducted to ascertain potential risk factors for the recurrence in colorectal cancer patients. These factors encompass histopathological considerations, lifestyle choices, genetic factors, clinical characteristics, comorbidities, and anthropometric indices[5-8]. The carcinoembryonic antigen (CEA), a component of standard postoperative surveillance, is routinely monitored alongside chest and abdominal CT scans, colonoscopy, and other assessments. Its frequent measurement every 3-6 months also makes it suitable for time-dependent risk assessment for recurrence[4,9]. Identifying risk factors necessitates quantitative analyses, often requiring the utilization of models. In line with the recent trend of utilizing machine learning models for medical data analysis, there has been a steady increase in studies aimed at predicting the survival of colorectal cancer patients. These studies employ machine learning models and utilize variables obtained from patient data[10-12].

Machine learning-based approaches offer distinct advantages in the assimilation and evaluation of clinical data. In comparison to traditional statistical methods, they provide greater flexibility and scalability, making them suitable for applications in diagnosis, treatment, and survival prediction. And, in contrast to conventional statistics, where the model is determined by researchers, machine learning-based analysis does not assume a specific model but rather constructs the model based on available data and algorithms[13]. This makes it well-suited for big data research and offers advantages in deriving optimal predictive results from complex interactions among various variables, and in retrospective studies based on electronic medical records, when the sample size is large, it becomes possible to overcome the limitations of insufficient or inaccurate individual patient information. Another notable advantage of machine learning-based approaches is their capacity to analyze diverse data types and integrate them into disease predictions[14]. Given these advantages, machine learning models have been applied in numerous studies analyzing data from hundreds to thousands of patients[10,12,15,16], and we also employed such an analytical approach. Furthermore, there have been advancements in algorithms capable of handling censored data, and the opportunities to leverage these algorithms are on the rise[17]. In addition to employing machine learning for precise predictions, there is a growing interest in understanding how models can be interpreted and how their prediction results can be explained. This field is known as interpretable machine learning (IML), which is defined as the process of extracting pertinent knowledge from a machine learning model, encompassing relationships inherent in the data or acquired by the model itself[18].

While several studies have employed machine learning to predict the prognosis of colorectal cancer patients, there has been limited development of machine learning models capable of concurrently predicting both recurrence and survival within the same patient group, particularly for Korean patients. Therefore, the primary objective of this study is to bridge this gap by developing machine learning-based models capable of predicting both recurrence and survival outcomes for stage I-III colorectal cancer patients who have undergone surgical interventions. Furthermore, using the developed time-to-event models, we endeavored to create an R Shiny application designed to facilitate the straightforward assessment of recurrence and survival probabilities for individual patients.

## II. MATERIALS AND METHODS

### 1. Patients and data collection

Our dataset included a cohort of 4,020 Korean patients who were diagnosed with stage I-III colorectal cancer and underwent surgical procedures at Severance Hospital in Seoul, South Korea. These patients were subject to follow-up evaluations, with intervals of at least three months during the initial two years post-surgery, followed by six-month intervals for the subsequent three years, and annual check-ups thereafter. CEA concentrations were routinely measured during each outpatient visit, abdominopelvic CT scans were conducted every six months for the first five years, and chest CT scans were performed annually following surgery.

This study was conducted in compliance with the principles outlined in the Declaration of Helsinki and received approval from the Institutional Review Board of Severance Hospital, Yonsei University College of Medicine (Seoul, South Korea). Patient consent was waived as this was a retrospective study.

### 2. Variables

Clinical features, including clinical stage (AJCC), histologic grade, tumor size, demographic information, operation-related details, and CEA concentrations, were systematically gathered from all patients as potential predictors, as these factors are recognized as risk factors linked to colorectal cancer prognosis[4,19,20]. The selection of these variables involved a screening process that considered p-values computed through statistical methods such as the log-rank test and Cox regression models. Additionally, during the selection of candidate predictors, the inter-variable correlation and the presence of multicollinearity were taken into account.

We also introduced a predictor named ‘Slope’, which was derived from the CEA samples, with the aim of improving the model's predictive performance. The values were estimated through linear regression using CEA samples from each individual patient, as shown in **Figure 1**. CEA samples prior to recurrence, or samples measured up to approximately 1 year in patients who did not experience recurrence, were used to create this variable.

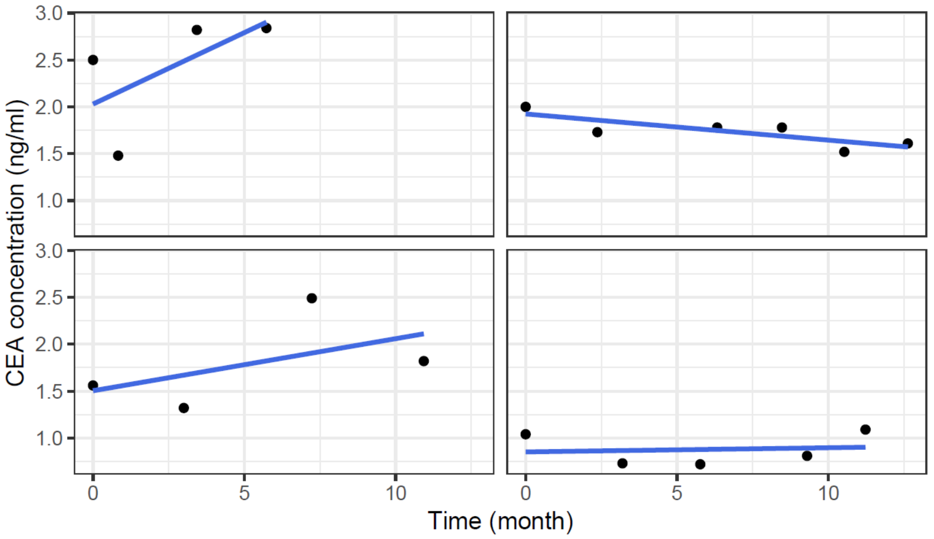


**Figure 1. Example of linear regression for CEA samples.** Black dots indicate observations, and blue lines represent linear regression lines.

### 3. Outcomes

The endpoints encompassed the presence of recurrence and the 5-year survival following surgical intervention, which served as criteria for classification models. Furthermore, time-to-event models incorporated the time to recurrence and overall survival as additional parameters. Recurrence was diagnosed through clinical and radiological examinations.

## 4. Statistical and machine learning-based modeling methods

The statistical analysis methods employed in this study comprised basic descriptive statistics, nonparametric techniques, such as the Kaplan-Meier method, and a semi-parametric approach using the Cox proportional hazard model. These procedures encompassed the development of machine learning-based models, the re-evaluation of predictive models, and the selection of relevant variables.

Machine learning-based predictive models were constructed by partitioning patient data into training (75%) and validation (25%) sets, selected randomly from individuals with records of recurrence or survival. For classification models, a variety of algorithms were employed, including logistic regression, support vector machine, decision tree, random forest, gradient boost, XGboost, Light-GBM, and CatBoost, in order to forecast the presence of recurrence and 5-year survival. Random survival forests, an ensemble method well-suited for handling right-censored data commonly applied for medical datasets, were utilized to predict recurrence and survival probabilities over time[17].

Feature selection was performed by assessing the significance of variables derived from each algorithm in the screened predictors and evaluating their impact on predictive performance. All models underwent hyperparameter tuning, accomplished through a 5-fold cross-validation process within the training set, utilizing the R caret package.

### 5. Performance assessment

Based on the developed models, predictive performance was evaluated using a validation dataset. The assessment involved calculating the Area Under the Receiver Operating Characteristic Curve (AUROC) as a measure of performance. Given the substantial class imbalance in the endpoints, performance was assessed by evaluating balanced accuracy. Additionally, precision, recall, and Kappa statistics were computed as performance indicators.

### 6. Software

R software (Version 4.2.2), in conjunction with its associated packages, was employed for all data analysis, the development of machine learning-based models, and the creation of R Shiny applications as outlined in the methods.

## III. RESULTS

### 1. Patient information

A total of 2,318 patients were included in the machine learning-based prediction models. Out of the initial 4,020 patients, 563 individuals who were unable to compute the variable 'Slope', from CEA samples and 1,139 patients who did not possess all of the potential predictors were excluded from the data set. The overall recurrence rate was approximately 13.5%, with the majority of recurrences occurring within 3 years following surgery. In contrast, the 5-year mortality rate stood at about 7.9%, which was lower than the recurrence rate. All variables related to CEA exhibited significant differences ($p < 0.05$) in accordance with recurrence and survival. Detailed demographic information about the patients, including potential predictors, is presented in **Tables 1 and 2**.

**Table 1.** The demographic information related to patients based on recurrence

| Recurrence | NO (N=2004) | | YES (N=314) | | P$^{*(T)}$ |
|---|---|---|---|---|---|
| | Continuous variables: mean (SD) | | | | |
| Age (year) | 61.2 (11.2) | | 60.8 (11.7) | | 0.581 |
| Pre-operative CEA (ng/ml) | 5.0 (9.1) | | 6.6 (12.5) | | 0.025 |
| Post-operative CEA (ng/ml) | 2.1 (3.8) | | 3.2 (8.3) | | 0.023 |
| Number of infiltrated nodes | 1.0 (2.2) | | 3.2 (5.3) | | <0.001 |
| Number of excised nodes | 21.8 (14.0) | | 21.7 (15.3) | | 0.921 |
| Slope of CEA (ng/ml/month) | 0.0 (0.4) | | 0.8 (4.2) | | 0.001 |
| | Categorical variables: number (%) | | | | P$^{*(C)}$ |
| T stage | T1 | 137 (6.8) | T1 | 4 (1.3) | <0.001 |
| | T2 | 397 (19.8) | T2 | 25 (8.0) | |
| | T3 | 1328 (66.3) | T3 | 234 (74.5) | |
| | T4 | 142 (7.1) | T4 | 51 (16.2) | |
| Tumor location | Right-sided | 497 (24.8) | Right-sided | 62 (19.7) | <0.001 |
| | Left-sided | 788 (39.3) | Left-sided | 100 (31.8) | |
| | Rectum, Anus | 719 (35.9) | Rectum, Anus | 152 (48.4) | |
| Histologic grade | Well | 269 (13.4) | Well | 29 (9.2) | 0.009 |
| | Moderately | 1597 (79.7) | Moderate | 251 (79.9) | |
| | Poorly | 138 (6.9) | Poorly | 34 (10.8) | |
| Lymphovascular invasion | No | 1615 (80.6) | No | 198 (63.1) | <0.001 |
| | Yes | 389 (19.4) | Yes | 116 (36.9) | |
| Microsatellite instability | MSI-H | 200 (10.0) | MSI-H | 15 (4.8) | 0.004 |
| | MSI-L, MSS | 1804 (90.0) | MSI-L, MSS | 299 (95.2) | |

*P-values were calculated by t-test$^{(T)}$ or Chi-squared test$^{(C)}$; SD, standard deviation.

**Table 2.** The demographic information related to patients based on 5-year survival

| Recurrence | NO (N=2004) | | YES (N=314) | | $P^{*(T)}$ |
|---|---|---|---|---|---|
| | Continuous variables: mean (SD) | | | | |
| Age (year) | 60.9 (11.2) | | 64.2 (12.5) | | 0.001 |
| Pre-operative CEA (ng/ml) | 5.1 (9.3) | | 7.2 (12.9) | | 0.028 |
| Post-operative CEA (ng/ml) | 2.2 (4.3) | | 3.4 (7.9) | | 0.046 |
| Number of infiltrated nodes | 1.1 (2.5) | | 3.4 (5.1) | | <0.001 |
| Number of excised nodes | 21.6 (13.9) | | 23.5 (17.0) | | 0.15 |
| Slope of CEA (ng/ml/month) | 0.0 (1.0) | | 1.0 (4.4) | | 0.003 |
| | Categorical variables: number (%) | | | | $P^{*(C)}$ |
| T stage | T1 | 138 (6.5) | T1 | 3 (1.6) | <0.001 |
| | T2 | 405 (19.0) | T2 | 17 (9.2) | |
| | T3 | 1434 (67.2) | T3 | 128 (69.6) | |
| | T4 | 157 (7.4) | T4 | 36 (19.6) | |
| Tumor location | Right-sided | 517 (24.2) | Right-sided | 42 (22.8) | 0.075 |
| | Left-sided | 829 (38.8) | Left-sided | 59 (32.1) | |
| | Rectum, Anus | 788 (36.9) | Rectum, Anus | 83 (45.1) | |
| Histologic grade | Well | 279 (13.1) | Well | 19 (10.3) | <0.001 |
| | Moderately | 1710 (80.1) | Moderate | 138 (75.0) | |
| | Poorly | 145 (6.8) | Poorly | 27 (14.7) | |
| Lymphovascular invasion | No | 1704 (79.9) | No | 109 (59.2) | <0.001 |
| | Yes | 430 (20.1) | Yes | 75 (40.8) | |
| Microsatellite instability | MSI-H | 203 (9.5) | MSI-H | 12 (6.5) | 0.226 |
| | MSI-L, MSS | 1931 (90.5) | MSI-L, MSS | 172 (93.5) | |

*P-values were calculated by t-test$^{(T)}$ or Chi-squared test$^{(C)}$; SD, standard deviation.

## 2. Statistical approaches

To facilitate statistical analysis, such as the log-rank test, the continuous variable was divided into two groups using the R maxstat package. For the two endpoints, distinct cut-points were applied to continuous variables for grouping. As an example, the values for CEA before and after the operation were determined as 3 and 3 ng/ml for recurrence, respectively, while they were set at 10 and 2.5 ng/ml for 5-year survival. In the log-rank test, all variables were found to be statistically significant ($p < 0.05$) for recurrence, with the exception of age ($p = 0.05$). For 5-year survival, most variables also demonstrated statistical significance ($p < 0.05$), except for tumor location ($p = 0.33$) and microsatellite instability ($p = 0.12$). The Kaplan-Meier plots, along with p-values calculated from the log-rank test and the cut-points for the continuous variables, can be found in **Appendices**.

The odds ratios, estimated through logistic regression, are presented in **Tables 3 and 4**. The analysis revealed that the number of infiltrated nodes, the number of excised nodes, Slope, tumor location, and T stage were significant factors in both analyses. Among these variables, the 'Slope' exhibited the highest odds ratios of 14.7 and 18.51 for recurrence and 5-year survival, respectively. However, age and lymphovascular invasion were found to be significant only in the context of 5-year survival. The results of the Cox proportional hazard model are presented in **Tables 5 and 6**. Post-operative CEA, number of infiltrated nodes, Slope, T stage, tumor location, lymphovascular invasion, and microsatellite instability were found to be statistically significant ($p<0.05$) in multivariate analysis for recurrence. Conversely, age and histologic grade were also statistically significant ($p<0.05$), while postoperative CEA and MSI no longer exhibited statistical significance in the 5-year survival analysis.

**Table 3.** Odds ratios for recurrence estimated from logistic regression

| Recurrence | | OR (95% CI) | |
|---|---|---|---|
| Variables | | Univariate | Multivariate |
| Age (year) | ≤ 74 | - | - |
| | > 74 | 1.26 (0.86-1.80, p=0.227) | 0.79 (0.50-1.24, p=0.318) |
| Pre-operative CEA (ng/ml) | ≤ 3 | - | - |
| | > 3 | 1.36 (1.07-1.72, p=0.013) | 0.91 (0.66-1.25, p=0.575) |
| Post-operative CEA (ng/ml) | ≤ 3 | - | - |
| | > 3 | 1.41 (1.03-1.90, p=0.028) | 1.22 (0.79-1.86, p=0.363) |
| Number of infiltrated nodes | ≤ 1 | - | - |
| | > 1 | 4.01 (3.14-5.13, p<0.001) | 2.74 (2.00-3.76, p<0.001) |
| Number of excised nodes | ≤ 10 | - | - |
| | > 10 | 0.68 (0.51-0.92, p=0.010) | 0.50 (0.35-0.73, p<0.001) |
| Slope of CEA (ng/ml/month) | ≤ 0.05 | - | - |
| | > 0.05 | 14.70 (11.18-19.40, p<0.001) | 14.87 (10.98-20.28, p<0.001) |
| T stage | T1 | - | - |
| | T2 | 2.16 (0.82-7.42, p=0.160) | 2.19 (0.73-8.35, p=0.200) |
| | T3 | 6.04 (2.52-19.77, p<0.001) | 6.64 (2.39-24.30, p=0.001) |
| | T4 | 12.30 (4.86-41.49, p<0.001) | 10.01 (3.30-38.68, p<0.001) |
| Tumor location | Right-sided | - | - |
| | Left-sided | 1.02 (0.73-1.43, p=0.920) | 0.87 (0.58-1.31, p=0.497) |
| | Rectum, Anus | 1.69 (1.24-2.34, p=0.001) | 1.85 (1.24-2.78, p=0.003) |
| Histologic grade | Well | - | - |
| | Moderately | 1.46 (0.99-2.23, p=0.068) | 1.00 (0.63-1.65, p=0.991) |
| | Poorly | 2.29 (1.34-3.93, p=0.003) | 1.21 (0.61-2.39, p=0.578) |
| Lymphovascular invasion | No | - | - |
| | Yes | 2.43 (1.88-3.13, p<0.001) | 1.35 (0.97-1.87, p=0.075) |
| Microsatellite instability | MSI-H | - | - |
| | MSI-L, MSS | 2.21 (1.33-3.95, p=0.004) | 1.84 (0.99-3.63, p=0.064) |

CI, confidence interval.

**Table 4.** Odds ratios for 5-year survival estimated from logistic regression

| Recurrence | | OR (95% CI) | |
|---|---|---|---|
| Variables | | Univariate | Multivariate |
| Age (year) | ≤ 66 | - | - |
| | > 66 | 2.13 (1.57-2.89, p<0.001) | 2.04 (1.42-2.93, p<0.001) |
| Pre-operative CEA (ng/ml) | ≤ 10 | - | - |
| | > 10 | 1.93 (1.28-2.84, p=0.001) | 1.37 (0.76-2.43, p=0.294) |
| Post-operative CEA (ng/ml) | ≤ 2.5 | - | - |
| | > 2.5 | 1.82 (1.30-2.51, p<0.001) | 1.15 (0.71-1.84, p=0.568) |
| Number of infiltrated nodes | ≤ 1 | - | - |
| | > 1 | 3.64 (2.68-4.95, p<0.001) | 1.96 (1.32-2.88, p=0.001) |
| Number of excised nodes | ≤ 35 | - | - |
| | > 35 | 1.53 (1.01-2.26, p=0.038) | 1.99 (1.17-3.33, p=0.010) |
| Slope of CEA (ng/ml/month) | ≤ 0.085 | - | - |
| | > 0.085 | 18.51 (13.25-26.01, p<0.001) | 14.30 (9.99-20.58, p<0.001) |
| T stage | T1 | - | - |
| | T2 | 1.93 (0.64-8.36, p=0.299) | 2.01 (0.60-9.34, p=0.303) |
| | T3 | 4.11 (1.53-16.80, p=0.017) | 2.85 (0.94-12.53, p=0.103) |
| | T4 | 10.55 (3.70-44.41, p<0.001) | 5.48 (1.63-25.58, p=0.013) |
| Tumor location | Right-sided | - | - |
| | Left-sided | 0.88 (0.58-1.33, p=0.528) | 1.21 (0.72-2.05, p=0.484) |
| | Rectum, Anus | 1.30 (0.89-1.93, p=0.189) | 2.11 (1.28-3.55, p=0.004) |
| Histologic grade | Well | - | - |
| | Moderately | 1.19 (0.74-2.01, p=0.502) | 0.74 (0.43-1.35, p=0.309) |
| | Poorly | 2.73 (1.48-5.15, p=0.001) | 1.20 (0.56-2.60, p=0.638) |
| Lymphovascular invasion | No | - | - |
| | Yes | 2.73 (1.99-3.72, p<0.001) | 1.61 (1.08-2.39, p=0.018) |
| Microsatellite instability | MSI-H | - | - |
| | MSI-L, MSS | 1.51 (0.86-2.90, p=0.182) | 1.17 (0.59-2.50, p=0.669) |

CI, confidence interval.

**Table 5.** Cox proportional hazard model for recurrence

| Recurrence | | HR (95% CI) | |
|---|---|---|---|
| Variables | | Univariate | Multivariate |
| Continuous variables: mean (SD) | | | |
| Age (year) | | 1.00 (0.99-1.01, p=0.957) | 1.00 (0.99-1.01, p=0.920) |
| Pre-operative CEA (ng/ml) | | 1.02 (1.01-1.03, p<0.001) | 1.00 (0.99-1.01, p=0.715) |
| Post-operative CEA (ng/ml) | | 1.03 (1.02-1.04, p<0.001) | 1.02 (1.01-1.04, p=0.012) |
| Number of infiltrated nodes | | 1.08 (1.07-1.09, p<0.001) | 1.07 (1.05-1.08, p<0.001) |
| Number of excised nodes | | 1.00 (0.99-1.01, p=0.892) | 0.99 (0.98-1.00, p=0.237) |
| Slope of CEA (ng/ml/month) | | 1.10 (1.08-1.13, p<0.001) | 1.07 (1.05-1.10, p<0.001) |
| Categorical variables: number (%) | | | |
| T stage | T1 | - | |
| | T2 | 2.21 (0.77-6.36, p=0.140) | 2.13 (0.74-6.15, p=0.161) |
| | T3 | 5.85 (2.18-15.72, p<0.001) | 5.38 (1.98-14.60, p=0.001) |
| | T4 | 12.56 (4.54-34.76, p<0.001) | 9.68 (3.42-27.44, p<0.001) |
| Tumor location | Right-sided | - | |
| | Left-sided | 1.02 (0.74-1.40, p=0.915) | 0.93 (0.65-1.32, p=0.667) |
| | Rectum, Anus | 1.51 (1.12-2.03, p=0.006) | 1.52 (1.09-2.14, p=0.015) |
| Histologic grade | Well | - | |
| | Moderately | 1.50 (1.02-2.20, p=0.038) | 1.08 (0.73-1.59, p=0.709) |
| | Poorly | 2.34 (1.42-3.84, p=0.001) | 1.55 (0.93-2.60, p=0.092) |
| Lymphovascular invasion | No | - | |
| | Yes | 2.30 (1.83-2.89, p<0.001) | 1.55 (1.22-1.96, p<0.001) |
| Microsatellite instability | MSI-H | - | |
| | MSI-L, MSS | 2.21 (1.31-3.71, p=0.003) | 1.78 (1.03-3.08, p=0.038) |

CI, confidence interval; SD, standard deviation.

**Table 6.** Cox proportional hazard model for 5-year survival

| Recurrence | | HR (95% CI) | |
|---|---|---|---|
| Variables | | Univariate | Multivariate |
| Continuous variables: mean (SD) | | | |
| Age (year) | | 1.03 (1.02-1.05, $p<0.001$) | 1.04 (1.02-1.05, $p<0.001$) |
| Pre-operative CEA (ng/ml) | | 1.02 (1.01-1.03, $p<0.001$) | 1.01 (1.00-1.03, $p=0.136$) |
| Post-operative CEA (ng/ml) | | 1.03 (1.01-1.04, $p<0.001$) | 1.01 (0.99-1.04, $p=0.335$) |
| Number of infiltrated nodes | | 1.07 (1.06-1.09, $p<0.001$) | 1.06 (1.04-1.08, $p<0.001$) |
| Number of excised nodes | | 1.01 (1.00-1.02, $p=0.073$) | 1.01 (1.00-1.02, $p=0.089$) |
| Slope of CEA (ng/ml/month) | | 1.11 (1.09-1.14, $p<0.001$) | 1.09 (1.07-1.12, $p<0.001$) |
| Categorical variables: number (%) | | | |
| T stage | T1 | - | |
| | T2 | 2.08 (0.61-7.11, $p=0.241$) | 2.08 (0.61-7.11, $p=0.241$) |
| | T3 | 4.14 (1.32-13.01, $p=0.015$) | 4.14 (1.32-13.01, $p=0.015$) |
| | T4 | 11.47 (3.53-37.25, $p<0.001$) | 11.47 (3.53-37.25, $p<0.001$) |
| Tumor location | Right-sided | - | - |
| | Left-sided | 0.89 (0.60-1.32, $p=0.563$) | 0.89 (0.60-1.32, $p=0.563$) |
| | Rectum, Anus | 1.15 (0.79-1.66, $p=0.474$) | 1.15 (0.79-1.66, $p=0.474$) |
| Histologic grade | Well | - | |
| | Moderately | 1.27 (0.79-2.05, $p=0.331$) | 1.27 (0.79-2.05, $p=0.331$) |
| | Poorly | 2.86 (1.59-5.15, $p<0.001$) | 2.86 (1.59-5.15, $p<0.001$) |
| Lymphovascular invasion | No | - | |
| | Yes | 2.52 (1.88-3.38, $p<0.001$) | 2.52 (1.88-3.38, $p<0.001$) |
| Microsatellite instability | MSI-H | - | |
| | MSI-L, MSS | 1.58 (0.88-2.83, $p=0.127$) | 1.58 (0.88-2.83, $p=0.127$) |

CI, confidence interval; SD, standard deviation.

### 3. Machine learning-based approaches

Before employing the machine learning-based approach, the data was randomly split into a 3:1 ratio to validate the developed models, and no significant differences ($p > 0.05$) were observed between the data sets with respect to potential predictors. The recurrence rate and mortality rate, serving as endpoints, were 13.7% and 8.1% in the training set, and 13.2% and 7.4% in the validation set, respectively. Further details are provided in **Table 7**.

**Table 7.** Demographic information of patients in the training and validation sets

| | Train set (N= 1748) | | Validation set (N=570) | | P$^{*(T)}$ |
|---|---|---|---|---|---|
| | | Continuous variables: mean (SD) | | | |
| Age (year) | 61.2 (11.4) | | 61.1 (11.0) | | 0.79 |
| Pre-operative CEA (ng/ml) | 5.2 (9.9) | | 5.2 (9.0) | | 0.836 |
| Post-operative CEA (ng/ml) | 2.3 (4.7) | | 2.2 (4.7) | | 0.818 |
| Number of infiltrated nodes | 1.3 (2.7) | | 1.3 (3.3) | | 0.789 |
| Number of excised nodes | 22.0 (14.6) | | 21.1 (12.6) | | 0.15 |
| Slope of CEA (ng/ml/month) | 0.1 (1.3) | | 0.2 (2.2) | | 0.333 |
| | | Categorical variables: number (%) | | | P*$^{(c)}$ |
| Cancer recurrence | No | 1509 (86.3) | No | 495 (86.8) | 0.809 |
| | Yes | 239 (13.7) | Yes | 75 (13.2) | |
| 5-year survival | No | 142 (8.1) | No | 42 (7.4) | 0.624 |
| | Yes | 1606 (91.9) | Yes | 528 (92.6) | |
| T stage | T1 | 108 (6.2) | T1 | 33 (5.8) | 0.648 |
| | T2 | 313 (17.9) | T2 | 109 (19.1) | |
| | T3 | 1175 (67.2) | T3 | 387 (67.9) | |
| | T4 | 152 (8.7) | T4 | 41 (7.2) | |
| Location of primary tumor | Right-sided | 426 (24.4) | Right-sided | 133 (23.3) | 0.083 |
| | Left-sided | 687 (39.3) | Left-sided | 201 (35.3) | |
| | Rectum, Anus | 635 (36.3) | Rectum, Anus | 236 (41.4) | |
| Histologic grade | Well | 232 (13.3) | Well | 66 (11.6) | 0.444 |
| | Moderately | 1383 (79.1) | Moderate | 465 (81.6) | |
| | Poorly | 133 (7.6) | Poorly | 39 (6.8) | |
| Lymphovascular invasion | No | 1362 (77.9) | No | 451 (79.1) | 0.584 |
| | Yes | 386 (22.1) | Yes | 119 (20.9) | |
| Microsatellite instability | MSI-H | 163 (9.3) | MSI-H | 52 (9.1) | 0.951 |
| | MSI-L, MSS | 1585 (90.7) | MSI-L, MSS | 518 (90.9) | |

*P-values were calculated by t-test$^{(T)}$ or Chi-squared test$^{(C)}$

### A. Classification prediction models

The relative importance of the selected variables, as calculated from each classification model, is illustrated in **Figures 2 and 3** for recurrence and 5-year survival, respectively. It's important to note that the support vector machine, due to its poor performance, and the logistic regression model, as previously demonstrated in the statistical analysis, were excluded from these figures. In the recurrence prediction models, a total of 7 variables were included, with 'Slope' identified as the most important variable, possessing the highest average score of 0.45. In contrast, for the 5-year survival prediction models, 'Slope' emerged as the most crucial variable as well. However, unlike the recurrence prediction model, 'lymphovascular invasion' was also incorporated into the model, leading to a total of 8 selected variables.

Performance indicators obtained from all the models are provided in **Table 8**. Excluding the logistic regression model, the average AUROC value was 0.90 for recurrence and 0.88 for 5-year survival, respectively. The performance of the models was similar when considering the selected variables. Additionally, the prediction models yielded kappa values that generally fell within an acceptable range, ranging from 0.27 to 0.56. Due to the limitations of the data characterized by a highly imbalanced distribution of endpoints, precision was calculated to be relatively low. **Figures 4 and 5** display the receiver operating characteristic (ROC) curves for four models, which exhibited slightly better performance compared to the other models.

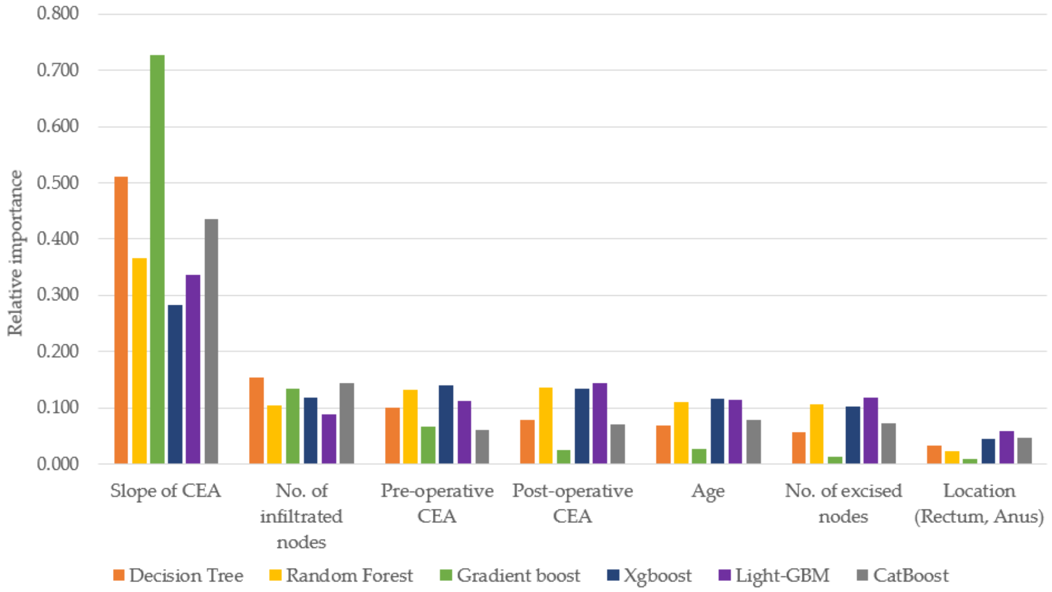


**Figure 2. Relative variable importance calculated from recurrence prediction models.**

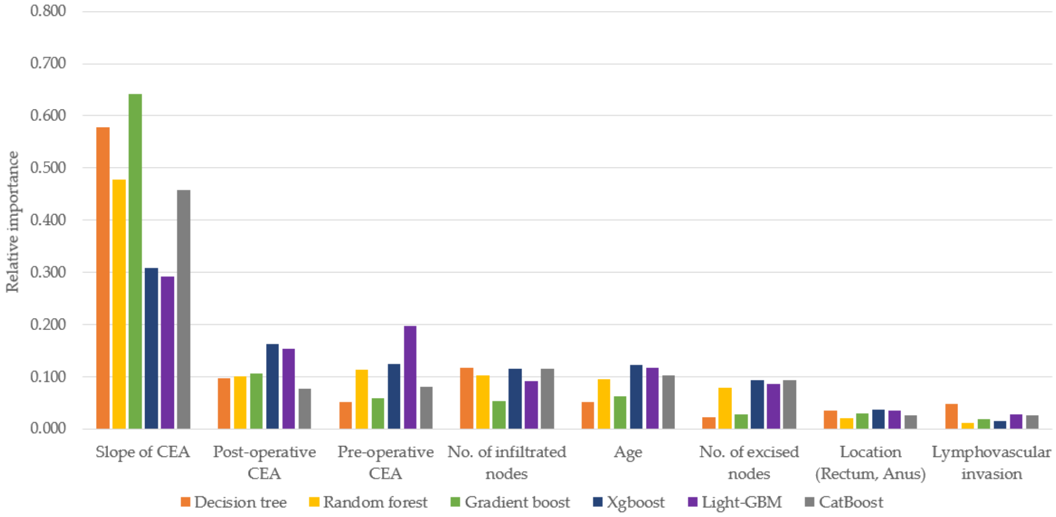


**Figure 3. Relative variable importance calculated from 5-year survival prediction models.**

**Table 8.** Performance of all classification prediction models

| Recurrence prediction models | | | | | |
|---|---|---|---|---|---|
| Model | Accuracy | Recall | Precision | Kappa | AUROC |
| Logistic regression | 0.78 | 0.77 | 0.36 | 0.38 | 0.85 |
| Decision tree | 0.77 | 0.59 | 0.65 | 0.56 | 0.87 |
| Random forest | 0.82 | 0.83 | 0.39 | 0.43 | 0.90 |
| Gradient boost | 0.81 | 0.83 | 0.37 | 0.40 | 0.89 |
| XGboost | 0.82 | 0.81 | 0.43 | 0.47 | 0.91 |
| Light-GBM | 0.83 | 0.84 | 0.42 | 0.47 | 0.91 |
| CatBoost | 0.83 | 0.83 | 0.44 | 0.49 | 0.92 |
| 5-year survival prediction models | | | | | |
| Model | Accuracy | Recall | Precision | Kappa | AUROC |
| Logistic regression | 0.68 | 0.52 | 0.20 | 0.20 | 0.75 |
| Decision tree | 0.82 | 0.79 | 0.30 | 0.36 | 0.87 |
| Random forest | 0.83 | 0.83 | 0.28 | 0.34 | 0.88 |
| Gradient boost | 0.81 | 0.81 | 0.27 | 0.33 | 0.86 |
| XGboost | 0.82 | 0.81 | 0.23 | 0.27 | 0.89 |
| Light-GBM | 0.79 | 0.81 | 0.23 | 0.27 | 0.89 |
| CatBoost | 0.82 | 0.83 | 0.25 | 0.30 | 0.89 |

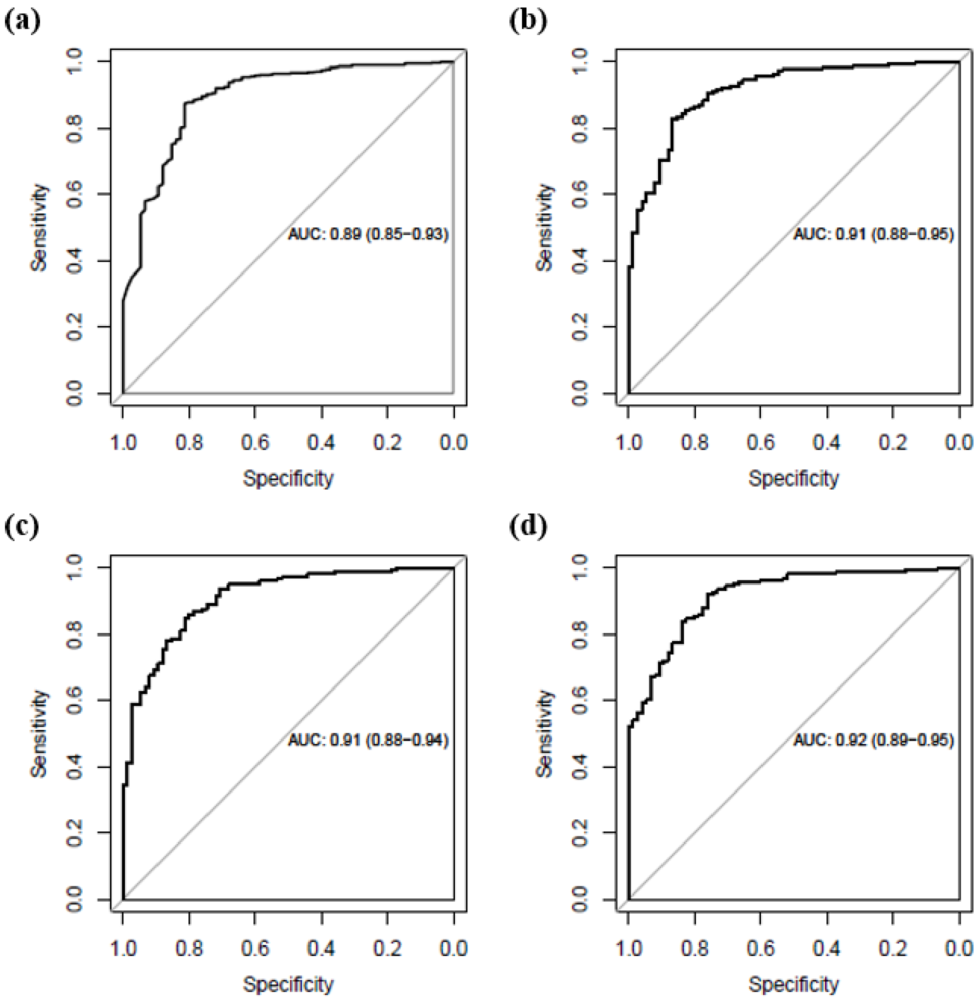


**Figure 4. Receiver operating characteristic (ROC) curves obtained from recurrence prediction models; (a) random forest, (b) XGboost, (c) Light-GBM, and (d) CatBoost.**

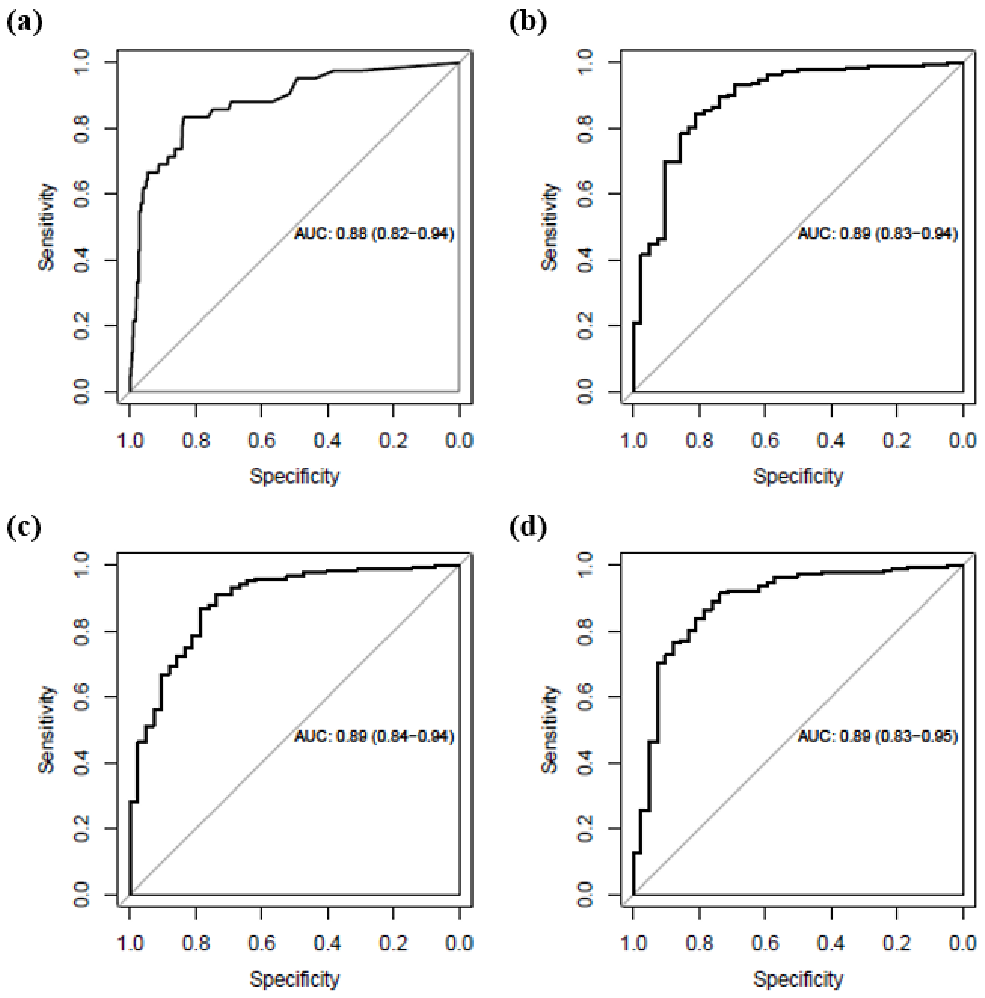


**Figure 5. Receiver operating characteristic (ROC) curves obtained from 5-year survival prediction models; (a) random forest, (b) XGboost, (c) Light-GBM, and (d) CatBoost.**

### B. Time-to-event models

Two models were developed using the R randomForestSRC package for predicting recurrence and survival probabilities over time. **Figures 6 and 7** describe the variable importance for both models and the marginal effects on all the selected variables. Six variables were selected for inclusion in both models, and the 'Slope' demonstrated relative influences exceeding 0.4, indicating its substantial contributions, similar to what was observed in the classification models. Furthermore, the continuous ranked probability score (CRPS) consistently remained below 0.1 throughout the designated target time in both models, as depicted in **Figure 8**. In contrast to the survival model, which was assessed over a 5-year timeframe, the recurrence model was examined up to 3 years, as this period encompassed the majority of recurrence events in our dataset. Regarding performance indicators, in the recurrence model, the AUROC was 0.90, while the accuracy, recall, precision, and kappa values stood at 0.84, 0.85, 0.43, and 0.48, respectively. Conversely, in the survival model, the AUROC was 0.89, and the accuracy, recall, precision, and kappa values were 0.78, 0.83, 0.20, and 0.22. Similar to the classification model, the prediction performance in the survival model was lower compared to the recurrence model. Based on the two developed models, **Figures 9 and 10** present the survival plots for 570 individuals in the validation dataset. Additionally, we selected three patients each who experienced recurrence or survival and three patients who did not, and visualized the results.

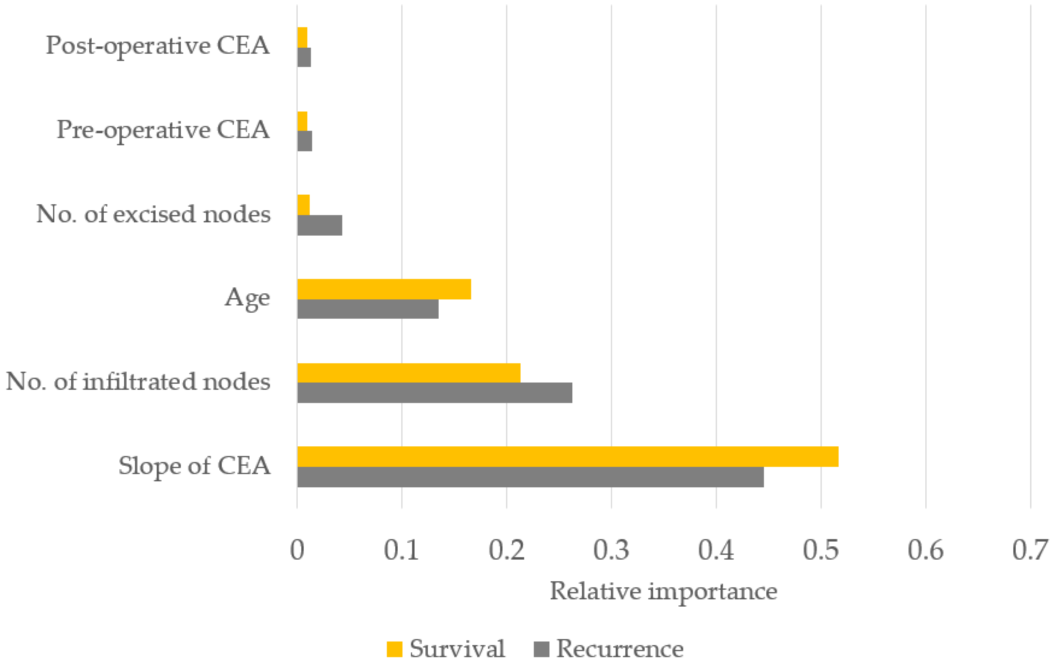


**Figure 6. Relative variable importance calculated from two time-to-event models.**

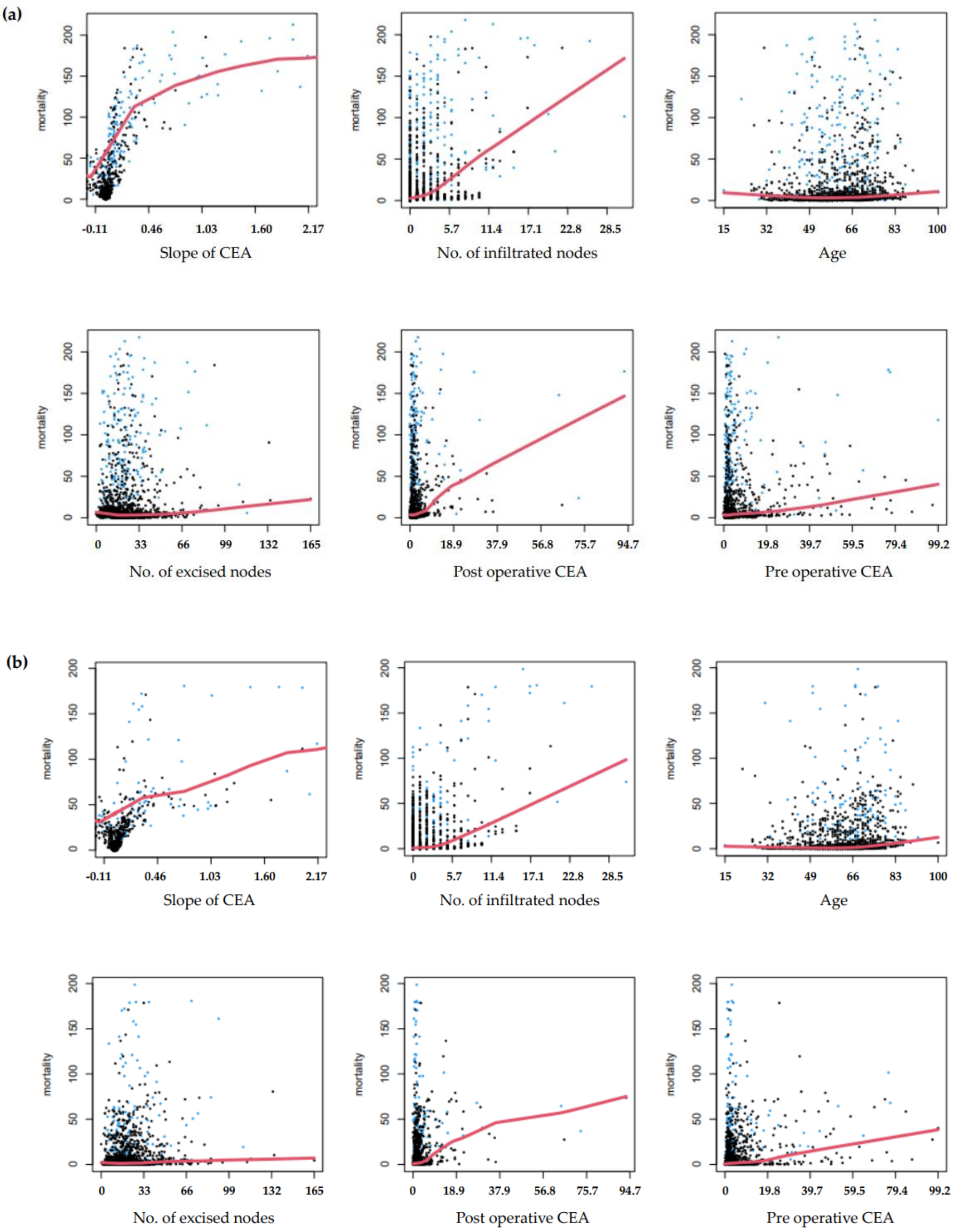

(a)
mortality
0
50
100
150
200
-0.11
0.46
1.03
1.60
2.17
Slope of CEA
0
5.7
11.4
17.1
22.8
28.5
No. of infiltrated nodes
15
32
49
66
83
100
Age
0
33
66
99
132
165
No. of excised nodes
0
18.9
37.9
56.8
75.7
94.7
Post operative CEA
0
19.8
39.7
59.5
79.4
99.2
Pre operative CEA
(b)
mortality
-0.11
0.46
1.03
1.60
2.17
Slope of CEA
No. of infiltrated nodes
Age
No. of excised nodes
Post operative CEA
Pre operative CEA

**Figure 7. The marginal effects of all the selected variables calculated from (a) the recurrence model and (b) the survival model is shown.** In the graphs, black dots represent observed data, while blue dots represent censored data.

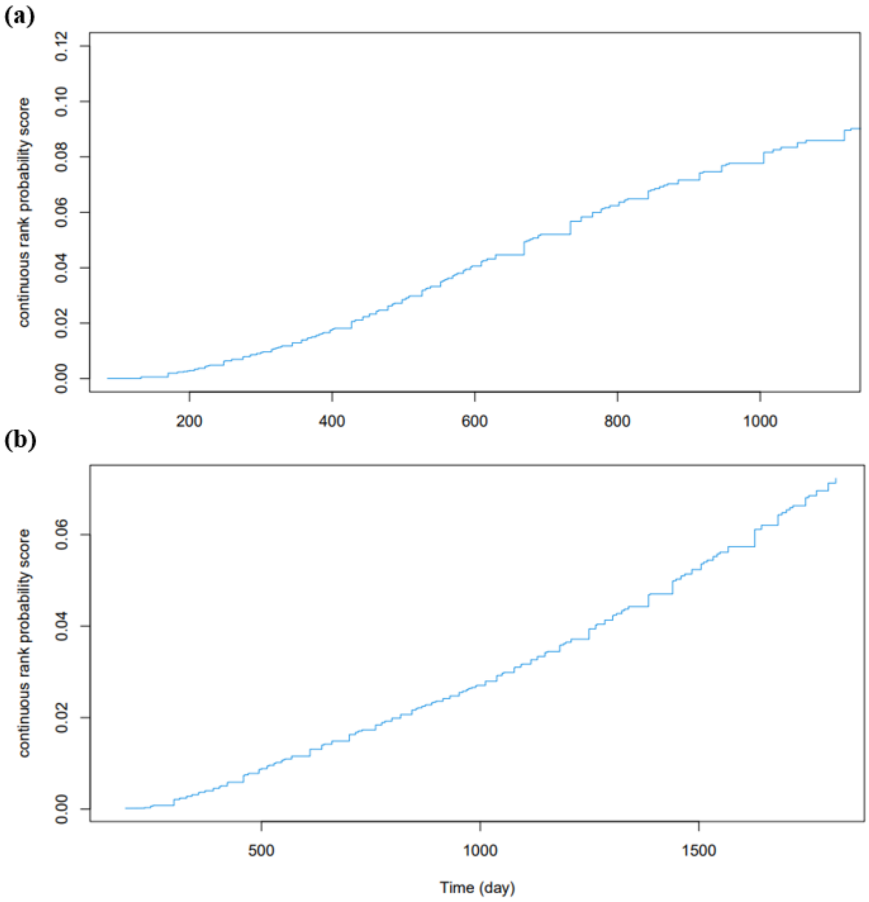


**Figure 8. Continuous rank probability score over time; (a) recurrence, (b) survival.**

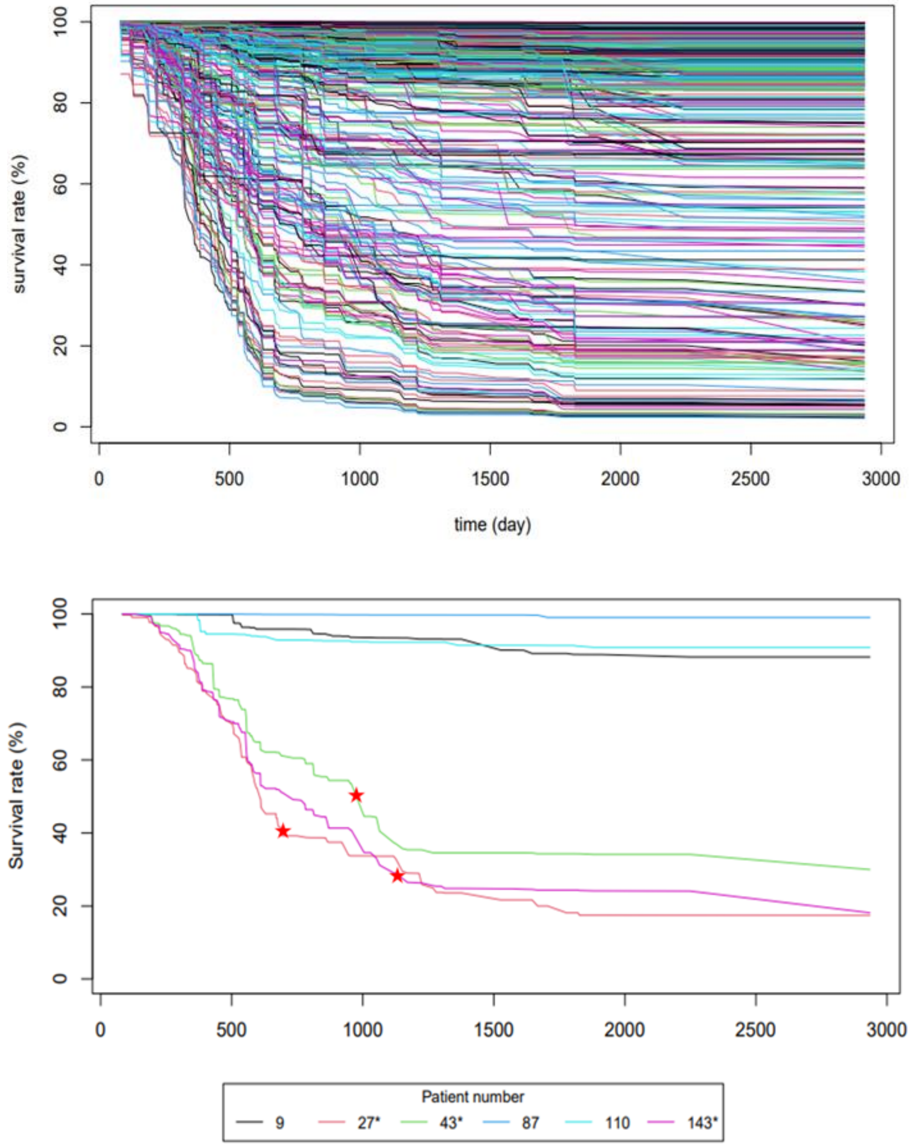


**Figure 9. Simulation results for the recurrence rate in a validation dataset (top) and for randomly sampled patients (bottom).** In the visualizations, an asterisk (*) denotes a patient with observed recurrence, and red stars indicate the same.

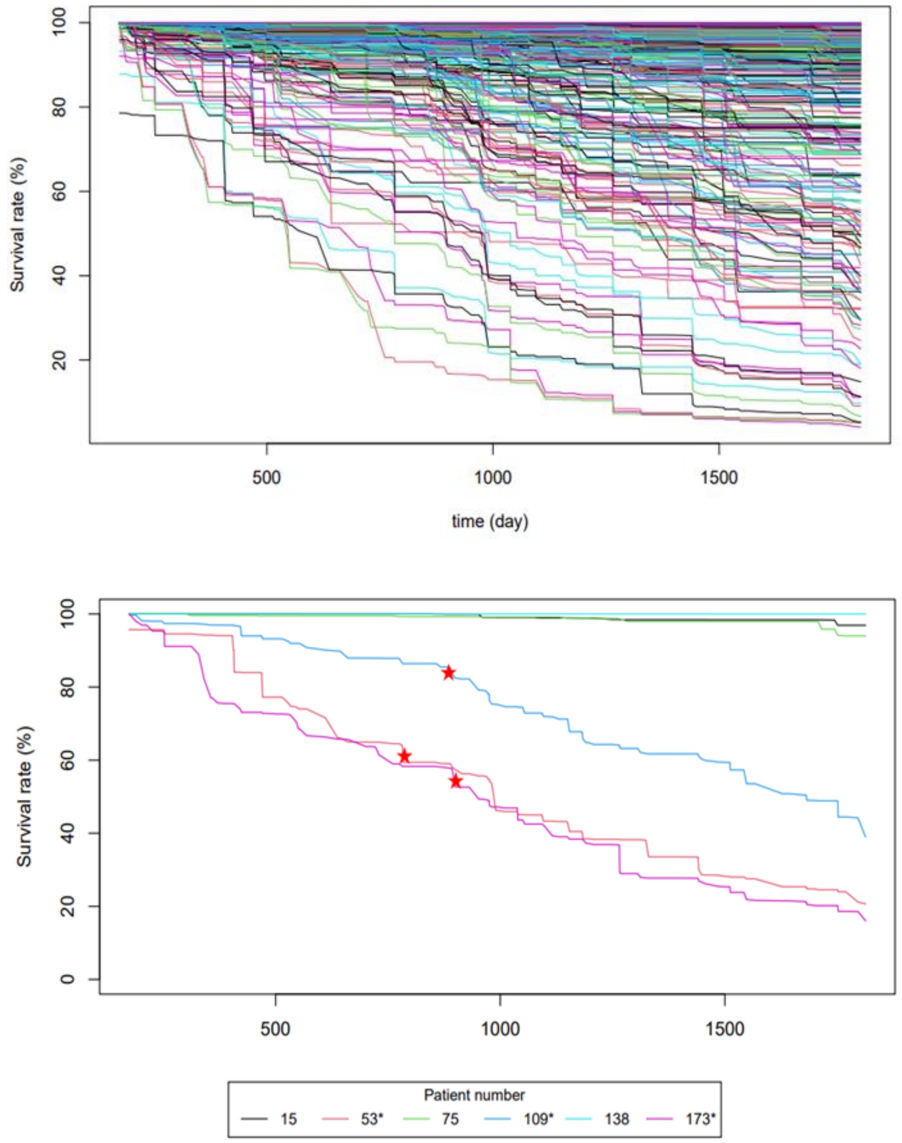
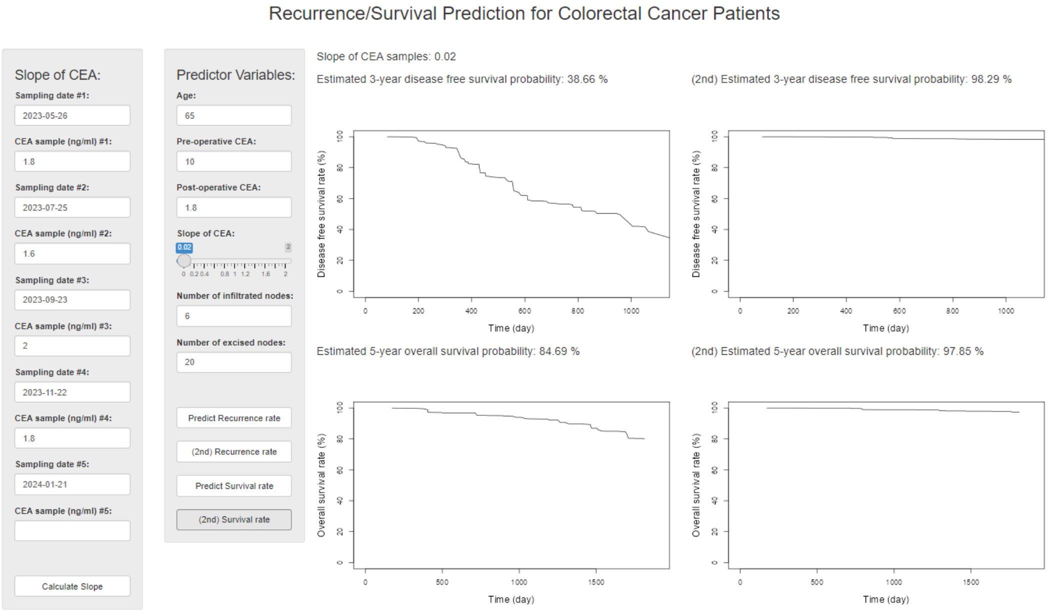


**Figure 10. Simulation results for the survival rate in a validation dataset (top) and for randomly sampled patients (bottom).** In the visualizations, an asterisk (*) denotes a patient with observed recurrence, and red stars indicate the same.

### C. R Shiny

The web-based application has been developed based on two time-to-event models, allowing users to visualize a patient's recurrence (disease free survival) and overall survival rate over time by inputting six predictor variables into the model. Additionally, users can obtain the slope generated from linear regression by providing CEA concentrations and their measurement dates, for the convenience of users. The application's user interface is displayed in **Figure 11**, providing a straightforward way to forecast outcomes for specific individual patients.



Recurrence/Survival Prediction for Colorectal Cancer Patients

Slope of CEA:

Sampling date #1: 2023-05-26
CEA sample (ng/ml) #1: 1.8
Sampling date #2: 2023-07-25
CEA sample (ng/ml) #2: 1.6
Sampling date #3: 2023-09-23
CEA sample (ng/ml) #3: 2
Sampling date #4: 2023-11-22
CEA sample (ng/ml) #4: 1.8
Sampling date #5: 2024-01-21
CEA sample (ng/ml) #5:

Predictor Variables:

Age: 65
Pre-operative CEA: 10
Post-operative CEA: 1.8
Slope of CEA: 0.02
Number of infiltrated nodes: 6
Number of excised nodes: 20

Slope of CEA samples: 0.02

Estimated 3-year disease free survival probability: 38.66 %

(2nd) Estimated 3-year disease free survival probability: 98.29 %

Estimated 5-year overall survival probability: 84.69 %

(2nd) Estimated 5-year overall survival probability: 97.85 %

**Figure 11. R Shiny application for predicting the prognosis of colorectal cancer patients.**

# IV. DISCUSSION

Machine learning-based models have been developed to predict the recurrence and survival of Korean colorectal cancer patients who underwent surgery, utilizing medical records. These models demonstrated strong performance, achieving AUROC values ranging from 0.87 to 0.92 for recurrence and 0.87 to 0.89 for survival, taking into consideration the selected variables influencing model performance. Furthermore, time-to-event models have also been developed to estimate the likelihood of recurrence and survival as time progresses, and their performance was 0.90 and 0.89 based on AUROC values.

Recently, there has been a growing interest in medical studies employing machine learning models. Consequently, numerous significant studies have been conducted in the field of colorectal cancer. A broad spectrum of research endeavors is underway, covering diverse aspects of this disease, including diagnosis, medical imaging, treatment, and prognosis. Previous studies have endeavored to forecast the survival of colorectal cancer patients[21,22]. In contrast to our study, a previous investigation employed immune genes instead of medical records. Their random survival forest model exhibited a 5-year survival prediction performance, achieving a Concordance index value of 0.818[21]. In a study involving Brazilian patients, various clinical features, including clinical staging, presence of recurrence, year of diagnosis, and surgery, were applied, demonstrating a survival performance with an AUC value of 0.86[22]. This study also exhibited commendable predictive performance; however, disparities in crucial variables were observed, likely attributable to variations in the target patient population.

Efforts have also been made to predict cancer recurrence. Disease-free survival was forecasted by incorporating patient information and a TAS score, which was determined based on tumor size, circumferential involvement, and tumor differentiation. The highest average AUC value achieved was 0.82 through a random forest model[10]. Notably, advanced age and a high lymph node ratio emerged as the most influential variables for predicting recurrence in patients with stage II–III colorectal cancer who underwent surgery. In Taiwanese studies, the projected AUC values from various models ranged from approximately 0.6 to 0.7[15]. Conversely, within our patient dataset, no substantial difference in tumor size ($p > 0.05$) was observed in relation to the presence of recurrence or survival. Consequently, it had minimal impact on model performance and was ultimately excluded as a predictor variable. A study conducted on Chinese patients in stage IV after surgery aimed to predict recurrence[16]. It identified chemotherapy, age, CEA levels, and anesthesia time as the most significant variables. The AUC value obtained in this study, reaching 0.761, showcased the highest performance in the gradient boost model.

Similar to our study, a South African study aimed to simultaneously predict recurrence and prognosis[12]. In addition to clinical features, this study incorporated numerous data points. Among the various models employed, the artificial neural network demonstrated the highest performance, yielding AUC values of 0.87 for recurrence and 0.81 for survival. Unlike prior studies where histology consistently emerged as the most pivotal variable in predicting survival, our study excluded histological findings or grades, as they had a minimal impact on model performance. This divergence is likely attributed to disparities between the target patient population and the applied predictors.

There was also a survival prediction study involving Korean patients as validation data, which confirmed significant differences between the characteristics of the Surveillance, Epidemiology, and End Results (SEER) data and Korean data[11]. Furthermore, the study revealed that prediction performance based on survival probability in the Light gradient boosting model was significantly superior to prediction based on the AJCC stage. The key variables identified as influential in this context were age, lymph node count, and tumor size. Therefore, research is needed to create a prediction model for Korean patients, and there have also been few cases of simultaneously developing a recurrence and survival prediction model in all populations.

One noteworthy aspect of the results is the inclusion of three CEA-related variables as predictors in each of the developed models. Carcinoembryonic antigen (CEA) has been widely used as a tumor marker since its initial description in 1965, and it is known to exhibit elevated levels in colorectal cancer patients[9,23]. However, it is important to note that CEA, while valuable, lacks specificity and can also increase in certain non-colorectal cancer conditions. Additionally, it may be measured as a normal value in early-stage colorectal cancer patients due to its limited sensitivity[24]. Given the continued challenge of early diagnosis, ongoing research is exploring alternatives such as long non-coding RNAs like PVT1, as well as molecular biomarkers like KRAS and MSI[25]. Nevertheless, due to its accessibility and cost-effectiveness, CEA is still recommended for postoperative surveillance alongside medical imaging in previous studies[4,26]. Therefore, its inclusion as a predictor in machine learning-based models holds an advantage, as it can be readily obtained in most clinical settings.

Thresholds were computed for colorectal cancer recurrence and survival, with the thresholds set at 3 ng/ml for postoperative CEA in the case of recurrence and 2.5 ng/ml for survival in our study. These values are consistent with those previously reported in the literature[9,27-29]. Furthermore, in an effort to enhance predictive performance, the postoperative CEA change rate was introduced as a novel predictor variable. Remarkably, this variable was established as the most critical predictor in all statistical analyses and machine learning-based models. It is established that CEA concentrations typically require approximately a month to return to within the normal range after surgery, with variations in both the time taken to reach the baseline and the extent of fluctuations observed[9]. Hence, given the challenge of incorporating the complete time profiles of CEA samples post-surgery, we opted to utilize the slope derived from linear regression. This approach allowed us to capture the overall change in CEA concentration while mitigating the impact of the final measurement. Certainly, it is important to acknowledge that this approach comes with a drawback: the inability to promptly assess a patient's prognosis immediately after surgery, as it necessitates time for sample collection to generate this derived variable. Nonetheless, the inclusion of this variable was deemed worthwhile due to its ability to enhance performance, resulting in an AUROC improvement of 0.15 or more compared to the best model created without the predictor, ‘Slope’. It also remains valuable, especially when considering that about 76% of recurrences occur after the first year following surgery.

We also compared the performance of repeatedly measured CEA concentrations after surgery and the rate of change obtained from linear regression to determine if they were clinically useful. The analysis was conducted using samples from the same period that were used to create the variable 'Slope’, ensuring a comparison under the same conditions. The cut-off value for CEA concentration was set at 5 ng/mL, which is the upper limit of the normal range. The cut-off value for 'Slope' was set at 0.045 ng/mL/month, based on the log-rank test. Among the 570 patients in the validation set, a total of 72 patients exceeded the normal range at least once, resulting in a sensitivity of 0.23 and a specificity of 0.89 in the prediction of recurrence using repeatedly measured CEA concentrations. While the specificity was high, the sensitivity was quite lower compared to values reported in previous studies for known colorectal cancer recurrences, typically around 0.4 or higher[24,30,31]. This discrepancy may be attributed to the limited observation period in this analysis. On the other hand, the sensitivity and specificity when using the rate of change obtained from linear regression were 0.67 and 0.92, respectively, demonstrating a higher predictive accuracy. Furthermore, among the 498 patients whose CEA concentration consistently remained within the normal range (less than 5 ng/mL), 58 experienced cancer recurrence. However, using the rate of change, recurrence could be predicted in 36 of those 58 patients. Conversely, recurrence was not observed in 55 out of the 72 patients who exceeded the normal range at least once. Using the rate of change, it was predicted that recurrence would not occur in 43 out of those 55 patients. Based on the analysis above, our data confirms that employing the rate of change obtained from linear regression is more effective in early-detecting recurrence in colorectal cancer patients than relying solely on measurements from repeatedly sampled after surgery.

Machine learning-based models also used the number of infiltrated lymph nodes, the number of excised lymph nodes, age, and tumor location as predictors. The influence of infiltrated lymph node counts on the prognosis of patients is well established, as in AJCC staging. Our time-to-event models also found that the predicted recurrence and mortality rate increased linearly with the number of infiltrated lymph nodes. Regarding the number of excised lymph nodes, there have been numerous discussions on the quantity of lymph nodes that should be examined after surgery[32-34]. Considering the marginal effect estimated from the time-to-event model for recurrence, although interactions with other variables such as clinical stage and the number of infiltrated lymph nodes should be taken into account when making interpretations, it appeared to be advantageous for the prognosis to examine about 15 lymph nodes. This is due to the fact that a greater number of excised lymph nodes was predicted to increase the risk of recurrence, whereas in cases with 15 or fewer resected lymph nodes, the lower the number of excised lymph nodes, the higher the risk of recurrence (**Figure 7**). Regarding age, recurrence and mortality rates generally increased with age. However, the prognosis was predicted to be poor in younger patients, which is presumed to be due to their poor clinical baseline. This point was consistent with national and international statistics[35,36]. Finally, patients with tumors in the rectum or anus were predicted to have a worse prognosis, which is consistent with findings from previous Japanese studies[37,38].

Recurrence was observed in approximately 13.5% of patients, with 84% of these cases occurring within the initial three years. These findings align closely with what has been reported in prior studies[39,40]. The trend in cumulative incidence rates exhibited similarity, although with lower overall recurrence rates among our patients. In addition, the five-year mortality rate was about 8%, which was lower than the recurrence rate. This aligns with the prognosis for non-metastatic colorectal cancer from previous literature[41]. However, such an imbalance in endpoints often leads to misconceptions when evaluating developed models. To mitigate this issue, we used 'balanced accuracy' as an evaluation metric, which considers weights based on the size of each class instead of 'accuracy,' thus providing a more reasonable assessment of the model's performance in the presence of imbalanced class distributions. We also presented Kappa values and ROC curves for a comprehensive evaluation. However, it's worth noting that, due to the nature of the data, the 'precision' value was relatively low compared to the values of other indicators. To address this problem, various resampling techniques are sometimes employed in the data preparation process for creating machine learning models[42]. However, in cases like ours with an extremely high class imbalance, undersampling can result in significant data loss, diluting the significance of analyzing big data, and simultaneously leading to decreased model performance. On the other hand, oversampling generally yields better results than undersampling, but it can introduce bias and overfitting issues, particularly in highly imbalanced data[42]. Furthermore, in our dataset, we did not observe any benefits from applying resampling techniques; therefore, we chose not to utilize them.

Machine learning-based models developed in this study demonstrated strong predictive performance for recurrence and survival in Korean colorectal cancer patients. This study also confirmed the advantage of using the CEA change rate estimated from linear regression for detecting recurrence and survival in patients. Furthermore, there is an advantage to developing predictive models using easily available routine clinical data in most clinical settings. However, this study also has limitations. External validation was not conducted for the developed models, and there are ethnic and genetic differences not only in the most important variable, CEA, but also in relation to the prognosis of colorectal cancer[43,44]. However, this also implies the need for developing prediction models for specific populations.

## V. CONCLUSION

Machine learning-based models were developed to predict survival and recurrence among Korean patients with stage I-III colorectal cancer who underwent surgery, using medical records. The classification models exhibited AUROC values of about 0.9, indicative of their strong predictive performance. The most important predictor was the created variable, Slope, followed by the number of infiltrated nodes, excised lymph node counts, age, CEA concentrations before and after surgery, and tumor location. And the time-to-event models developed using the random survival forest also exhibited good performance. In addition, an R Shiny application has been developed based on time-to-event models to facilitate the easy evaluation of the prognosis of colorectal cancer in individual patients.

## REFERENCES

1. OECD (2021), Health at a Glance 2021: OECD Indicators, OECD Publishing, Paris.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
3. Statistics Korea, Korean Statistical Information Service (KOSIS), Daejeon, Republic of Korea.
4. Steele SR, Chang GJ, Hendren S, Weiser M, Irani J, Buie WD, et al. Practice Guideline for the Surveillance of Patients After Curative Treatment of Colon and Rectal Cancer. Dis Colon Rectum 2015;58:713-25.
5. Xu W, He Y, Wang Y, Li X, Young J, Ioannidis JPA, et al. Risk factors and risk prediction models for colorectal cancer metastasis and recurrence: an umbrella review of systematic reviews and meta-analyses of observational studies. BMC Med 2020;18:172.
6. Osterman E, Glimelius B. Recurrence Risk After Up-to-Date Colon Cancer Staging, Surgery, and Pathology: Analysis of the Entire Swedish Population. Dis Colon Rectum 2018;61:1016-25.
7. Jaspan V, Lin K, Popov V. The impact of anthropometric parameters on colorectal cancer prognosis: A systematic review and meta-analysis. Crit Rev Oncol Hematol 2021;159:103232.
8. Boakye D, Rillmann B, Walter V, Jansen L, Hoffmeister M, Brenner H. Impact of comorbidity and frailty on prognosis in colorectal cancer patients: A systematic

review and meta-analysis. Cancer Treat Rev 2018;64:30-9.

9. Saito G, Sadahiro S, Kamata H, Miyakita H, Okada K, Tanaka A, et al. Monitoring of Serum Carcinoembryonic Antigen Levels after Curative Resection of Colon Cancer: Cutoff Values Determined according to Preoperative Levels Enhance the Diagnostic Accuracy for Recurrence. Oncology 2017;92:276-82.

10. Gupta P, Chiang SF, Sahoo PK, Mohapatra SK, You JF, Onthoni DD, et al. Prediction of Colon Cancer Stages and Survival Period with Machine Learning Approach. Cancers (Basel) 2019;11.

11. Osman MH, Mohamed RH, Sarhan HM, Park EJ, Baik SH, Lee KY, et al. Machine Learning Model for Predicting Postoperative Survival of Patients with Colorectal Cancer. Cancer Res Treat 2022;54:517-24.

12. Achilonu OJ, Fabian J, Bebington B, Singh E, Eijkemans MJC, Musenge E. Predicting Colorectal Cancer Recurrence and Patient Survival Using Supervised Machine Learning Approach: A South African Population-Based Study. Front Public Health 2021;9:694306.

13. Ley C, Martin RK, Pareek A, Groll A, Seil R, Tischer T. Machine learning and conventional statistics: making sense of the differences. Knee Surg Sports Traumatol Arthrosc 2022;30:753-7.

14. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20:e262-e73.

15. Chen PC, Yeh YM, Lin BW, Chan RH, Su PF, Liu YC, et al. A Prediction Model for Tumor Recurrence in Stage II-III Colorectal Cancer Patients: From a Machine Learning Model to Genomic Profiling. Biomedicines 2022;10.

16. Xu Y, Ju L, Tong J, Zhou CM, Yang JJ. Machine Learning Algorithms for

Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection. Sci Rep 2020;10:2519.

17. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Annals of Applied Statistics 2008;2:841-60.

18. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A 2019;116:22071-80.

19. Hoshino N, Hasegawa S, Hida K, Kawada K, Ganeko R, Sugihara K, et al. Nomogram for predicting recurrence in stage II colorectal cancer. Acta Oncol 2016;55:1414-7.

20. Ryuk JP, Choi GS, Park JS, Kim HJ, Park SY, Yoon GS, et al. Predictive factors and the prognosis of recurrence of colorectal cancer within 2 years after curative resection. Ann Surg Treat Res 2014;86:143-51.

21. Zhang Z, Huang L, Li J, Wang P. Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: a novel machine learning survival predictive system. BMC Bioinformatics 2022;23:124.

22. Buk Cardoso L, Cunha Parro V, Verzinhasse Peres S, Curado MP, Fernandes GA, Wunsch Filho V, et al. Machine learning for predicting survival of colorectal cancer patients. Sci Rep 2023;13:8874.

23. Thomson DM, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. Proc Natl Acad Sci U S A 1969;64:161-7.

24. Tan E, Gouvas N, Nicholls RJ, Ziprin P, Xynos E, Tekkis PP. Diagnostic precision of carcinoembryonic antigen in the detection of recurrence of colorectal cancer.

Surg Oncol 2009;18:15-24.

25. Ogunwobi OO, Mahmood F, Akingboye A. Biomarkers in Colorectal Cancer: Current Research and Future Prospects. Int J Mol Sci 2020;21.

26. Cervantes A, Adam R, Rosello S, Arnold D, Normanno N, Taieb J, et al. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 2023;34:10-32.

27. Thomas DS, Fourkala EO, Apostolidou S, Gunu R, Ryan A, Jacobs I, et al. Evaluation of serum CEA, CYFRA21-1 and CA125 for the early detection of colorectal cancer using longitudinal preclinical samples. Br J Cancer 2015;113:268-74.

28. Jeong S, Nam TK, Jeong JU, Kim SH, Kim K, Jang HS, et al. Postoperative carcinoembryonic antigen level has a prognostic value for distant metastasis and survival in rectal cancer patients who receive preoperative chemoradiotherapy and curative surgery: a retrospective multi-institutional analysis. Clin Exp Metastasis 2016;33:809-16.

29. Auclin E, Andre T, Taieb J, Banzi M, Van Laethem JL, Tabernero J, et al. Association of post-operative CEA with survival and oxaliplatin benefit in patients with stage II colon cancer: a post hoc analysis of the MOSAIC trial. Br J Cancer 2019;121:312-7.

30. Nicholson BD, Shinkins B, Pathiraja I, Roberts NW, James TJ, Mallett S, et al. Blood CEA levels for detecting recurrent colorectal cancer. Cochrane Database Syst Rev 2015;2015:CD011134.

31. Goldstein MJ, Mitchell EP. Carcinoembryonic antigen in the staging and follow-up of patients with colorectal cancer. Cancer Invest 2005;23:338-51.

32. Chang GJ, Rodriguez-Bigas MA, Skibber JM, Moyer VA. Lymph node evaluation and survival after curative resection of colon cancer: systematic review. J Natl Cancer Inst 2007;99:433-41.

33. Li Destri G, Di Carlo I, Scilletta R, Scilletta B, Puleo S. Colorectal cancer and lymph nodes: the obsession with the number 12. World J Gastroenterol 2014;20:1951-60.

34. Foo CC, Ku C, Wei R, Yip J, Tsang J, Chan TY, et al. How does lymph node yield affect survival outcomes of stage I and II colon cancer? World J Surg Oncol 2020;18:22.

35. Gabriel E, Attwood K, Al-Sukhni E, Erwin D, Boland P, Nurkin S. Age-related rates of colorectal cancer and the factors associated with overall survival. J Gastrointest Oncol 2018;9:96-110.

36. Park HC, Shin A, Kim BW, Jung KW, Won YJ, Oh JH, et al. Data on the characteristics and the survival of korean patients with colorectal cancer from the Korea central cancer registry. Ann Coloproctol 2013;29:144-9.

37. Hashiguchi Y, Muro K, Saito Y, Ito Y, Ajioka Y, Hamaguchi T, et al. Japanese Society for Cancer of the Colon and Rectum (JSCCR) guidelines 2019 for the treatment of colorectal cancer. Int J Clin Oncol 2020;25:1-42.

38. Shida D, Inoue M, Tanabe T, Moritani K, Tsukamoto S, Yamauchi S, et al. Prognostic impact of primary tumor location in Stage III colorectal cancer-right-sided colon versus left-sided colon versus rectum: a nationwide multicenter retrospective study. J Gastroenterol 2020;55:958-68.

39. Desch CE, Benson AB, Somerfield MR, Flynn PJ, Krause C, Loprinzi CL, et al. Colorectal cancer surveillance: 2005 update of an American Society of Clinical

Oncology practice guideline. J Clin Oncol 2005;23:8512-9.

40. Holmes AC, Riis AH, Erichsen R, Fedirko V, Ostenfeld EB, Vyberg M, et al. Descriptive characteristics of colon and rectal cancer recurrence in a Danish population-based study. Acta Oncol 2017;56:1111-9.

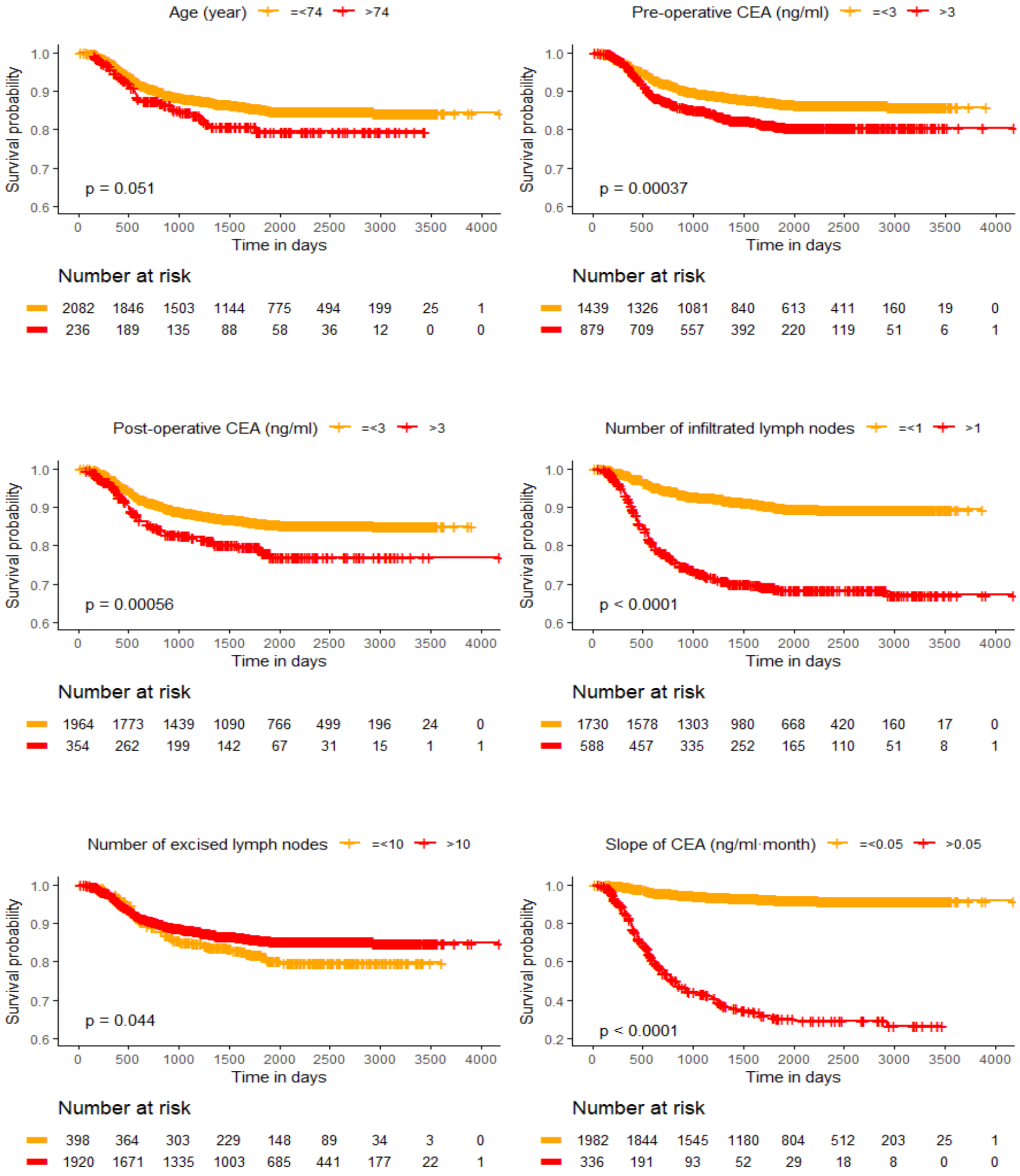
41. Li N, Lu B, Luo C, Cai J, Lu M, Zhang Y, et al. Incidence, mortality, survival, risk factor and screening of colorectal cancer: A comparison among China, Europe, and northern America. Cancer Lett 2021;522:255-68.

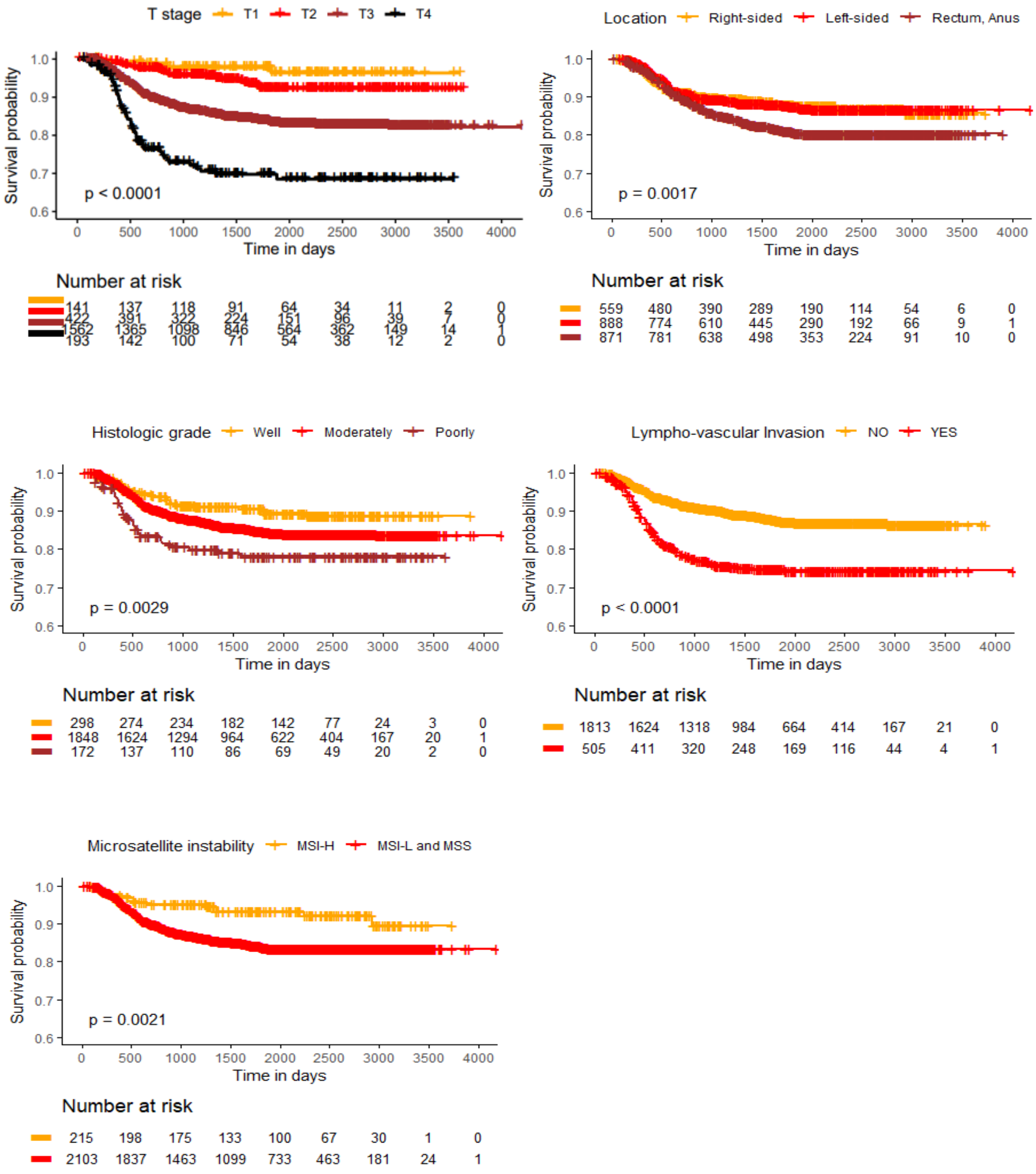
42. Werner de Vargas V, Schneider Aranda JA, Dos Santos Costa R, da Silva Pereira PR, Victoria Barbosa JL. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. Knowl Inf Syst 2023;65:31-57.

43. Liang Y, Tang W, Huang T, Gao Y, Tan A, Yang X, et al. Genetic variations affecting serum carcinoembryonic antigen levels and status of regional lymph nodes in patients with sporadic colorectal cancer from Southern China. PLoS One 2014;9:e97923.

44. Ollberding NJ, Nomura AM, Wilkens LR, Henderson BE, Kolonel LN. Racial/ethnic differences in colorectal cancer risk: the multiethnic cohort study. Int J Cancer 2011;129:1899-906.

# APPENDICES


Age (year) =<74 >74
Survival probability
1.0 0.9 0.8 0.7 0.6
p = 0.051
Time in days
0 500 1000 1500 2000 2500 3000 3500 4000
Number at risk
2082 1846 1503 1144 775 494 199 25 1
236 189 135 88 58 36 12 0 0
Pre-operative CEA (ng/ml) =<3 >3
Survival probability
1.0 0.9 0.8 0.7 0.6
p = 0.00037
Time in days
0 500 1000 1500 2000 2500 3000 3500 4000
Number at risk
1439 1326 1081 840 613 411 160 19 0
879 709 557 392 220 119 51 6 1
Post-operative CEA (ng/ml) =<3 >3
Survival probability
1.0 0.9 0.8 0.7 0.6
p = 0.00056
Time in days
0 500 1000 1500 2000 2500 3000 3500 4000
Number at risk
1964 1773 1439 1090 766 499 196 24 0
354 262 199 142 67 31 15 1 1
Number of infiltrated lymph nodes =<1 >1
Survival probability
1.0 0.9 0.8 0.7 0.6
p < 0.0001
Time in days
0 500 1000 1500 2000 2500 3000 3500 4000
Number at risk
1730 1578 1303 980 668 420 160 17 0
588 457 335 252 165 110 51 8 1
Number of excised lymph nodes =<10 >10
Survival probability
1.0 0.9 0.8 0.7 0.6
p = 0.044
Time in days
0 500 1000 1500 2000 2500 3000 3500 4000
Number at risk
398 364 303 229 148 89 34 3 0
1920 1671 1335 1003 685 441 177 22 1
Slope of CEA (ng/ml·month) =<0.05 >0.05
Survival probability
1.0 0.8 0.6 0.4 0.2
p < 0.0001
Time in days
0 500 1000 1500 2000 2500 3000 3500 4000
Number at risk
1982 1844 1545 1180 804 512 203 25 1
336 191 93 52 29 18 8 0 0



Number at risk

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 🟧 | 141 | 137 | 118 | 91 | 64 | 34 | 11 | 2 | 0 |
| 🟥 | 422 | 391 | 322 | 224 | 151 | 96 | 39 | 7 | 0 |
| 🟫 | 1562 | 1365 | 1098 | 846 | 564 | 362 | 149 | 14 | 1 |
| ⬛ | 193 | 142 | 100 | 71 | 54 | 38 | 12 | 2 | 0 |

Number at risk

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 🟧 | 559 | 480 | 390 | 289 | 190 | 114 | 54 | 6 | 0 |
| 🟥 | 888 | 774 | 610 | 445 | 290 | 192 | 66 | 9 | 1 |
| 🟫 | 871 | 781 | 638 | 498 | 353 | 224 | 91 | 10 | 0 |

Number at risk

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 🟧 | 298 | 274 | 234 | 182 | 142 | 77 | 24 | 3 | 0 |
| 🟥 | 1848 | 1624 | 1294 | 964 | 622 | 404 | 167 | 20 | 1 |
| 🟫 | 172 | 137 | 110 | 86 | 69 | 49 | 20 | 2 | 0 |

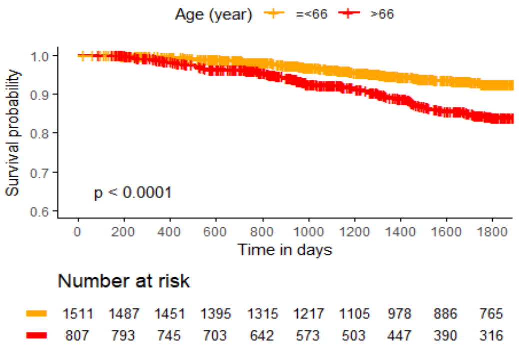
Number at risk

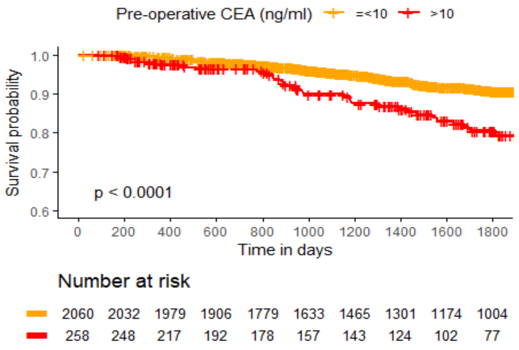
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 🟧 | 1813 | 1624 | 1318 | 984 | 664 | 414 | 167 | 21 | 0 |
| 🟥 | 505 | 411 | 320 | 248 | 169 | 116 | 44 | 4 | 1 |

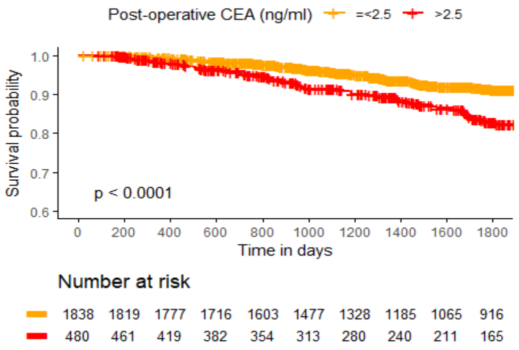
Number at risk

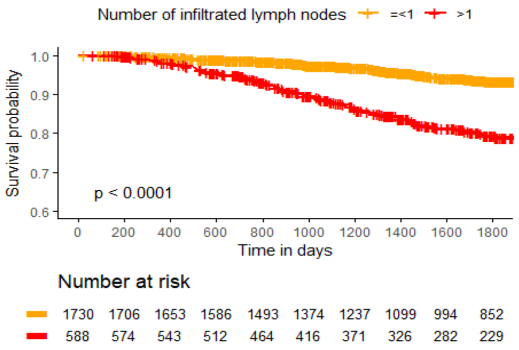
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 🟧 | 215 | 198 | 175 | 133 | 100 | 67 | 30 | 1 | 0 |
| 🟥 | 2103 | 1837 | 1463 | 1099 | 733 | 463 | 181 | 24 | 1 |

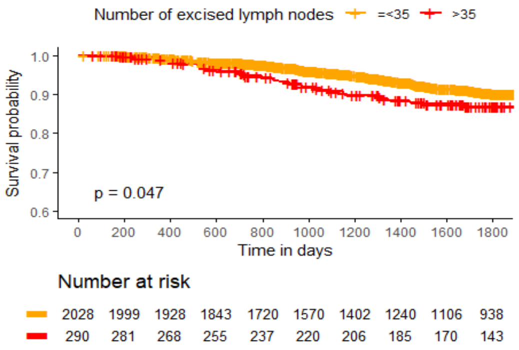
**Appendix 1**. Kaplan-Meier plots for potential predictors of recurrence.

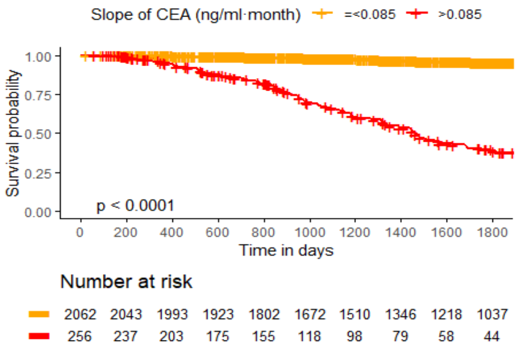

Age (year)
=<66
>66
Survival probability
1.0
0.9
0.8
0.7
0.6
p < 0.0001
0
200
400
600
800
1000
1200
1400
1600
1800
Time in days
Number at risk
1511 1487 1451 1395 1315 1217 1105 978 886 765
807 793 745 703 642 573 503 447 390 316

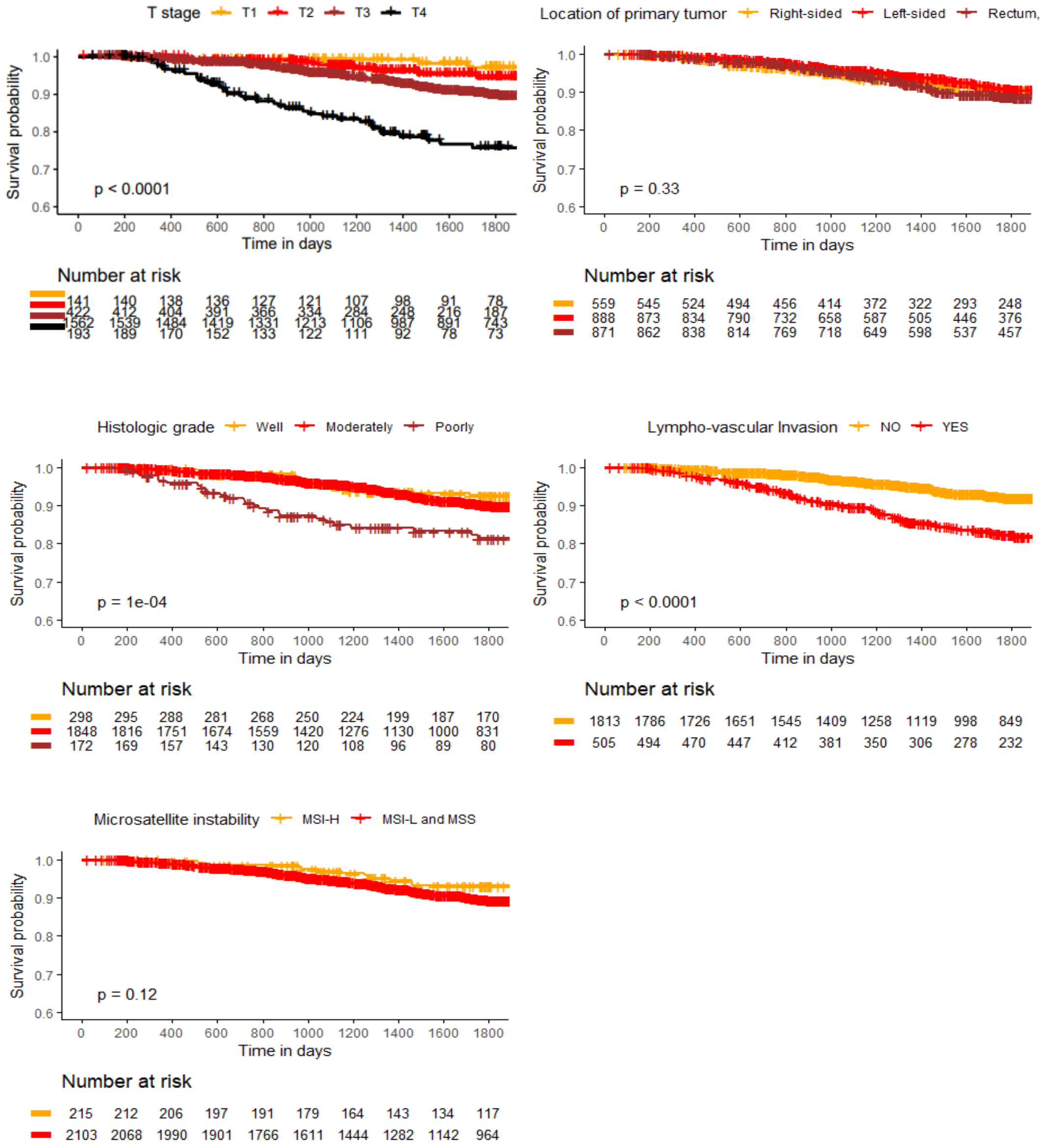

Pre-operative CEA (ng/ml)
=<10
>10
Survival probability
1.0
0.9
0.8
0.7
0.6
p < 0.0001
0
200
400
600
800
1000
1200
1400
1600
1800
Time in days
Number at risk
2060 2032 1979 1906 1779 1633 1465 1301 1174 1004
258 248 217 192 178 157 143 124 102 77


Post-operative CEA (ng/ml)
=<2.5
>2.5
Survival probability
1.0
0.9
0.8
0.7
0.6
p < 0.0001
0
200
400
600
800
1000
1200
1400
1600
1800
Time in days
Number at risk
1838 1819 1777 1716 1603 1477 1328 1185 1065 916
480 461 419 382 354 313 280 240 211 165


Number of infiltrated lymph nodes
=<1
>1
Survival probability
1.0
0.9
0.8
0.7
0.6
p < 0.0001
0
200
400
600
800
1000
1200
1400
1600
1800
Time in days
Number at risk
1730 1706 1653 1586 1493 1374 1237 1099 994 852
588 574 543 512 464 416 371 326 282 229


Number of excised lymph nodes
=<35
>35
Survival probability
1.0
0.9
0.8
0.7
0.6
p = 0.047
0
200
400
600
800
1000
1200
1400
1600
1800
Time in days
Number at risk
2028 1999 1928 1843 1720 1570 1402 1240 1106 938
290 281 268 255 237 220 206 185 170 143


Slope of CEA (ng/ml·month)
=<0.085
>0.085
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
200
400
600
800
1000
1200
1400
1600
1800
Time in days
Number at risk
2062 2043 1993 1923 1802 1672 1510 1346 1218 1037
256 237 203 175 155 118 98 79 58 44



**Appendix 2**. Kaplan-Meier plots for potential predictors of 5-year survival.

## ABSTRACT(IN KOREAN)

### 대장암 환자의 재발과 생존 예측을 위한 CEA 기반의 머신러닝 기법

<지도교수 박경수>

연세대학교 대학원 의과학과

윤석용

**목표**: 대장암은 선진국에서 암 관련 사망의 두 번째 주요 원인이며 적절한 치료를 받은 환자 중에도 일부는 암의 재발을 경험합니다. 재발 초기 진단은 환자의 예후를 향상시킨다는 것이 알려져 있습니다. 그럼에도 불구하고 현재 재발의 초기 감지를 가능하게 하는 비침습적 방법이 부족합니다. 이와 관련하여, 이 연구는 대장암 환자의 재발 및 생존을 조기에 예측하기 위한 방법을 개발하고자 하였습니다.

**자료 및 방법**: 세브란스병원에서 1-3 기 대장암으로 진단 후 수술 받은 4,020 명의 환자 자료를 기반으로 하고 있습니다. 각 환자로부터 재발 및 생존을 조기 예측하기 위한 잠재적 예측 변수로서 인구학적 정보 및 임상 특성, 수술 전 및 수술 후 CEA 농도, 침윤된 림프절 수, 절제된 림프절 수, 종양 위치 및 수술 시 나이 등이 수집되었습니다. 또한 다른 예측 변수인 'Slope' 라는 변수가 생성되었는데, 이것은 암태아성항원 (CEA)의 혈중 농도에서 파생된 것으로, 재발 전 또는 수술 후 약 1년까지의 CEA

샘플로부터 얻은 선형 회귀 기울기를 의미합니다. 예측 변수 중 어느 하나에 결측 값이 있는 환자는 제외되었습니다. 분석은 두 단계로 수행되었습니다. 첫 번째 단계에서, 재발 여부와 생존 여부를 예측하기 위한 분류 모델이 개발되었습니다. 두 번째 단계에서, 시간 의존적인 암의 재발 및 생존 확률을 예측하기 위한 모델이 개발되었습니다. 이 연구에는 유연성과 확장성이 있고 데이터만을 기반으로 구현할 수 있으며 특정 모델의 가정이 필요하지 않는 이점이 있는 머신 러닝이 사용되었습니다. 데이터 분석과 모델 개발은 R 소프트웨어 (버전 4.2.2) 및 패키지를 이용하여 수행되었습니다.

**결과**: 다양한 머신 러닝 알고리즘을 테스트하여 분류 모델을 개발했습니다. 이 알고리즘에는 로지스틱 회귀, 서포트 벡터 머신, 의사 결정 트리, 랜덤 포레스트, Gradient boost, XGboost, Light-GBM, 그리고 CatBoost 가 활용되었습니다. 재발 여부 예측 모델에서 ROC 곡선 아래 면적 (AUROC) 의 범위는 0.87-0.92, 생존 여부 예측 모델에서 AUROC 값의 범위는 0.87-0.89 였습니다. 이러한 모델 중에서 CatBoost 알고리즘이 적용된 모델이 약간 더 나은 성능을 나타냈습니다. Time-to-event 모델은 랜덤 서바이벌 포레스트 알고리즘을 이용하여 개발되었으며, 재발 모델에 대한 AUROC 값은 0.90 이며, 생존 모델에 대한 AUROC 값은 0.89 로 산출되었습니다. 모든 개발된 모델에서 새로 도입한 변수인 'Slope' 가 가장 중요한 예측 변수였습니다. 개발된 Time-to-event 모델에 기반하여, 개별 환자 수준의 예측을 용이하게 하기 위한 R Shiny 애플리케이션을 만들었습니다.

**결론**: 본 연구는 대장암에서 재발 및 생존 여부와 시간 의존적 재발 및 생존 확률을 조기에 예측하기 위한 CEA의 활용가능성을 확인했습니다. 환자의 예후 예측을 위해 개발된 모델은 성능이 좋았으며, 개발된 모델과 R Shiny 어플리케이션이 한국 대장암 환자의 예후 평가에 도움이 되길 기대합니다.

핵심되는 말 : 대장암, 머신러닝 기반 모델, 재발, 생존, 암태아성항원