A Comparison of Machine Learning Models for Colon Cancer Survival Estimation

Reuben Adatorwovor, Motolani E Ogunsanya, Bin Huang, Richard Charnigo, Olufunmilola Abraham

Submitted to: JMIR Cancer
on: March 09, 2025

Table of Contents

A Comparison of Machine Learning Models for Colon Cancer Survival Estimation

Reuben Adatorwovor[1] PhD; Motolani E Ogunsanya[2] DPH, PhD; Bin Huang[3] DrPH; Richard Charnigo[4] PhD; Olufunmilola Abraham[5] PhD, MS, BPharm

[1]Department of Biostatistics College of Public Health University of Kentucky Lexington US
[2]TSET Health Promotion Research Center Department of Family and Preventive Medicine University of Oklahoma Health Sciences Center Oklahoma City US
[3]Department of Internal medicine College of Medicine University of Kentucky Lexington US
[4]Dr. Bing Zhang Department of Statistics College of Arts and Sciences University of Kentucky Lexington US
[5]Department of Pharmacy Practice and Science College of Pharmacy University of Kentucky Lexington US

Corresponding Author:
Reuben Adatorwovor PhD
Department of Biostatistics
College of Public Health
University of Kentucky
760 Rose street
Suite 208H
Lexington
US

Abstract

Background: Colon cancer is a leading cause of cancer-related deaths worldwide, and survival outcomes are influenced by a variety of factors, including biological traits, treatment type, and patient characteristics. Traditional statistical models, such as Kaplan-Meier curves, have been widely used to estimate survival probabilities. However, these models often have difficulty handling complex interactions, covariates, and non-linear relationships between risk factors. Recently, machine learning techniques have emerged as promising tools for improving survival prediction by handling large covariates and capturing complex patterns.

Objective: This study compares several machine learning (ML) models to accurately estimate colon cancer survival by leveraging data from the Kentucky Cancer Registry (KCR). By identifying key risk factors, these analyses aims to improve risk stratification, treatment planning, and prognosis for overall colon cancer survival and within subgroups.

Methods: This retrospective study examines registry data to compare a variety of predictive modeling techniques, including Cox proportional hazards, accelerated failure time (AFT) models, random survival forests, Lasso, and Elastic Net, which were applied to the registry data to predict survival probabilities. The models were evaluated based on their predictive accuracy, feature importance, and ability to handle complex risk factors.

Results: Key covariates influencing survival outcomes, such as age, treatment type, positive nodes, tumor stage, smoking, and comorbidities, were identified as significant predictors of survival. The results highlight the strengths and limitations of each machine learning approach, with the random forest and Lasso models outperforming traditional methods in terms of prediction accuracy and identifying non-linear relationships.

Conclusions: This comparative analysis offers valuable insights for clinical decision-making and prognosis, highlighting the potential of machine learning to identify risk factors specific to different subgroups, ultimately advancing personalized care for colon cancer patients.

(JMIR Preprints 09/03/2025:72665)
DOI: https://doi.org/10.2196/preprints.72665

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Reuben Adatorwovor, PhD[1*], Motolani E. Ogunsanya, DPh, PhD[2,3], Bin Huang, DrPH[4], Richard Charnigo, PhD [1,5], Olufunmilola Abraham, PhD, MS, BPharm[6]

Affiliations:
1 Department of Biostatistics, College of Public Health, University of Kentucky
2 Department of Family and Preventive Medicine, University of Oklahoma Health Sciences, Oklahoma City, Oklahoma
3 TSET Health Promotion Research Center, University of Oklahoma Health Sciences, Oklahoma City, Oklahoma
4 Department of Internal medicine, College of Medicine, University of Kentucky
5 Dr. Bing Zhang Department of Statistics, College of Arts and Sciences, University of Kentucky
6 Department of Pharmacy Practice and Science, College of Pharmacy, University of Kentucky

*Corresponding author

A Comparison of Machine Learning Models for Colon Cancer Survival Estimation

ABSTRACT

Background:
Colon cancer is a leading cause of cancer-related deaths worldwide, and survival outcomes are influenced by a variety of factors, including risk factors, treatment type, and patient characteristics. Traditional statistical models, such as Kaplan-Meier curves, have been widely used to estimate survival probabilities. However, these models often have difficulty handling complex interactions, covariates, and non-linear relationships between risk factors. Recently, machine learning techniques have emerged as promising tools for improving survival prediction by handling large covariates and capturing complex patterns.

Objective:
This study compares several machine learning (ML) models to accurately estimate colon cancer survival by leveraging data from the Kentucky Cancer Registry (KCR). By identifying key risk factors, these analyses aim to improve risk stratification, treatment planning, and prognosis for overall colon cancer survival and within subgroups.

Method:
This retrospective study examines registry data to compare a variety of predictive modeling techniques, including Cox proportional hazards, accelerated failure time (AFT) models, random survival forests, Lasso, and Elastic Net, which were applied to the registry data to predict survival probabilities. The models were evaluated based on their predictive accuracy, feature importance, and ability to handle complex risk factors.

Results:
Key covariates influencing survival outcomes, such as age, treatment type, positive nodes, tumor

stage, smoking, and comorbidities, were identified as significant predictors of survival. The results highlight the strengths and limitations of each machine learning approach, with the random forest and Lasso models outperforming traditional methods in terms of prediction accuracy and identifying non-linear relationships.

Conclusions:
This comparative analysis offers valuable insights for clinical decision-making and prognosis, highlighting the potential of machine learning to identify risk factors specific to different subgroups, ultimately advancing personalized care for colon cancer patients.

Keywords: Colon cancer survival; colorectal cancer; Cox model; Elastic Net; Lasso; machine learning models; risk factors; random forest; survival estimation.

INTRODUCTION

Colon cancer, also known as colorectal cancer (CRC) is a leading cause of cancer-related deaths globally[1,2]. It affects the colon or rectum, which are parts of the large intestine. CRC typically develops from adenomatous polyps, abnormal growths in the colon lining that may take several years to become cancerous. Despite advancements in screening, early detection, and treatment, colon cancer remains a significant public health concern. Survival rates are influenced by several risk factors[1], including cancer stage at diagnosis, the presence of metastasis, and various individual and biological risk factors. Early-stage detection offers better survival outcomes, while more advanced stage detections are associated with significantly lower rates of survival. Given these variations in survival outcomes across different subgroups of the population, understanding the risk factors that influence prognosis is essential for improving patient care and survival outcomes.

Colon cancer survival rates in the United States have improved significantly over the past few decades, largely due to advancements in early detection, treatment, and enhanced screening protocols. According to the Centers for Disease Control and Prevention (CDC), the 5-year relative survival rate for colon cancer has increased from 50% in the 1970s to approximately 64% in recent years[3]. However, survival rates vary significantly by stage at diagnosis: localized colon cancers have a 5-year survival rate of approximately 90%, whereas late-stage diagnoses have survival rates as low as 14%[4]. These improvements were attributed to the growing implementation of colorectal cancer screening programs, such as colonoscopies, which allow for early detection and intervention. Despite these advances, disparities in survival persist based on factors like race, socioeconomic status, and access to healthcare, with non-Hispanic Black patients experiencing lower survival rates compared to non-Hispanic Whites[3].

These subgroup variations pose significant challenges for public health efforts. While national survival rates have improved, significant regional disparities remain. For example, in Kentucky, colon cancer survival rates are notably lower than the national average, highlighting key public health challenges in the region. The state's overall 5-year survival rate for colon cancer is approximately 61%, which is lower than the national average of 64%[3]. Additionally, Kentucky has one of the highest colorectal cancer mortality rates in the United States, partially due to high rates of smoking, obesity, and physical inactivity, all of which are modifiable risk factors for colon cancer[4]. Limited access to healthcare and early screening programs, particularly in rural or Appalachian

regions, contributes to delayed diagnoses and poorer survival outcomes[4,5]. Although Kentucky has implemented various cancer control initiatives, such as increasing access to screenings, these efforts have been less effective in rural areas, where healthcare resources remain scarce[6,7]. As a result, higher number of patients in Kentucky are diagnosed at later stages compared to the USA overall, which significantly impacts their chances of survival[8,9].

Risk Factors for Colon Cancer

Colon cancer is influenced by a combination of non-modifiable and modifiable risk factors. Age is one of the most significant non-modifiable risk factors, with most CRC cases diagnosed in individuals aged 60 and above[10]. The accumulation of genetic mutations and cellular damage over time is believed to contribute to the increased incidence of colon cancer in older populations[11]. Additionally, individuals with a family history of colon cancer particularly those with first-degree relatives (parents, siblings, or children) who have been diagnosed have a higher risk of developing the disease[12]. Certain inherited genetic syndromes, such as Familial Adenomatous Polyposis (FAP) and Lynch Syndrome, also increase the likelihood of early-onset CRC due to genetic mutations in mismatch repair genes[13,14].

Among modifiable risk factors, smoking and alcohol consumption significantly increase colon cancer risk. Carcinogens in tobacco smoke can damage colon cell DNA, while heavy alcohol intake contributes to the accumulation of acetaldehyde, a known carcinogen that promotes CRC development[15,16]. Studies show that individuals who both smoke and consume excessive alcohol face an even higher risk[17,18].

Physical inactivity, obesity, and chronic inflammatory conditions like Crohn's disease and ulcerative colitis are known risk factors for CRC[19]. Studies have shown that regular physical activity is associated with a reduced risk of CRC, as exercise enhances gastrointestinal motility, lowers systemic inflammation, and supports a healthy body weight[19,20]. Conversely, abdominal obesity increases CRC risk[21], as excess visceral fat triggers chronic inflammation and elevates hormone levels, such as insulin and leptin, which can stimulate cancer cell growth[22]. Research suggests that obesity-related colon cancer risk is particularly high for cancers located in the proximal colon[23]. Additionally, dietary habits, hormonal factors (especially in women), and screening practices play essential roles in CRC prevention and detection. Regular screening methods, including colonoscopy, sigmoidoscopy, and fecal occult blood tests, can detect precancerous polyps or early-stage cancers, significantly improving survival rates[24] with 5-year survival around 90%[25] for patients diagnosed at an early stage with localized cancer.

Survival Outcomes and Prognosis

The prognosis for colon cancer patients depends largely on the characteristics of the cancer at the time of diagnosis. These characteristics may include the stage, which describes the extent of cancer spread. The TNM staging system is commonly used to classify cancer based on tumor size (T), lymph node involvement (N), and the presence of metastasis (M). Patients diagnosed at Stages I and II typically have localized disease, with no lymph node involvement, and high survival rates (90-95%), when treated early[26]. In Stage III, cancer has spread to nearby lymph nodes, but not too distant organs, leading to moderate survival rates (40–70%), depending on treatment effectiveness[26]. Stage IV represents advanced disease, where cancer has metastasized to distant organs such as the liver or lungs, resulting in significantly lower survival rates (10–15%)[27]. However, newer treatment

approaches such as targeted therapies, immunotherapies, and palliative care have improved survival and quality of life for patients with metastatic disease.

Colon cancer continues to be a major global health challenge, with survival outcomes influenced by early detection, lifestyle modifications, and timely intervention. While modifiable risk factors such as diet, exercise, and smoking cessation play a critical role in cancer prevention, genetic predispositions and medical conditions like inflammatory bowel disease (IBD) further complicate individual risk profiles. Early screening remains the cornerstone of improving colon cancer survival, as timely detection dramatically increases the likelihood of successful treatment.

Despite progress in cancer surveillance and treatment, a major challenge remains: accurately predicting survival outcomes based on individual risk factors. Cancer survival surveillance using registry data is often complicated by incomplete or unreliable cause-of-death information, which limits the effectiveness of public health interventions and cancer control strategies. The aim of this paper is to explore how risk factors influence prognostic outcomes in cancer survival analysis using machine learning models. Specifically, we seek to answer two key questions: how do different risk factors impact cancer survival across various subgroups, and which risk factors are the strongest predictors in understanding survival outcomes within these subgroups?

METHODS

Data Source and Cohort Selection

The Kentucky Cancer Registry (KCR), a part of the Surveillance, Epidemiology, and End Results (SEER) Program and the National Program of Cancer Registries (NCR), provides comprehensive, population-based data on cancer incidence, survival, treatment, and patient demographic factors in Kentucky. KCR routinely linked the cancer incidence data to the Kentucky state mortality data and the National Death Index data to ensure most updated mortality information. Following approval from the Institutional Review Board (IRB #63067) and completion of the data use agreement, we extracted de-identified level II colon cancer data from the KCR for the most recently available period (January 1, 2010, to December 31st, 2022) for this retrospective study. The study included all non-institutionalized adult patients (aged ≥18 years) diagnosed with colon cancer during the study period. Patients with missing diagnosis, treatment, or staging data were excluded, ensuring data integrity.

Patient, Disease, and Treatment Factors

Patient demographic and risk factors such as age at diagnosis, sex, race, ethnicity, marital status, smoking status, cigarette pack years, and geographical residence (Appalachian or Non-Appalachian regions of Kentucky) were extracted from the KCR for patients diagnosed with colon cancer. Age at diagnosis was categorized as early onset (18- 60 years) or late (over 60 years). Due to the low representation of certain racial groups, patients who were neither White nor Black were categorized as “Other” to maintain statistical power. Ethnicity was classified into “Not of Spanish or Latino origin” and “Other” due to small sample sizes**.** Marital status was grouped into broader categories, with divorced and separated individuals combined into a single “Divorced or Separated” category, and those living with a partner or having unknown marital status grouped under “Unknown.”

Histological types[25] not matching “8140” (Adenocarcinoma NOS), “8210” (Papillary Adenocarcinoma, Papillary subtype), “8240” (Tubulovillous Adenocarcinoma), and “8480” (Mucinous Adenocarcinoma) were categorized as “Other/Unspecified.” Treatment records were

similarly streamlined based on the first course treatment: patients reporting no treatment, unknown treatment, or refusal were grouped under “No/Unknown or Refused Treatment.” Those undergoing surgery for primary or regional sites were listed as “Surgery (Primary and Regional),” while individuals receiving multi-modal treatment with surgery, chemotherapy, or radiation belonged to “Surgery + (Chemotherapy and/or Radiation).” Additional therapies were categorized as “Other.”

Colon cancer staging was often done using either the American Joint Committee on Cancer (AJCC) TNM system or the SEER Summary Stage system. The AJCC TNM staging is detailed, providing specific breakdowns of the tumor's size (T), regional lymph node involvement (N), and presence of distant metastasis (M). In contrast, we chose the SEER 2018 Summary Stage as it is a simplified system that categorizes cancer's spread into broader groups: In situ (confined to the origin site), localized (confined to the organ of origin without spread), regional (spread to nearby tissues or lymph nodes), distant (spread to distant organs), and unknown (extent of spread cannot be determined). Unlike the AJCC system, which offers a more granular, detailed assessment of cancer progression, the SEER Summary Stage aggregates these details into more general categories and more consistent across years, making it useful for large-scale epidemiological studies and population-based cancer survival tracking.

Insurance coverage was reclassified into four primary categories to streamline analysis. Patients with private or employer-based insurance were grouped under “Private Insurance.” Those covered by government-funded programs, including Medicare, Medicaid, TRICARE (Military), Workers’ Compensation, Veterans Affairs (VA), and the Indian Health Service (IHS), were classified as “Government-Related Programs.” Patients who were uninsured or self-paid were categorized as “Uninsured or Self-Pay,” while individuals with non-specific, unknown, or unreported insurance statuses were assigned to the “Other/Unknown Payers” group.

Statistical Analysis

We outlined key statistical approaches for analyzing the risk factors associated with colon cancer survival (from time of diagnosis until death or censored) in Kentucky adults. First, we provided a summary of colon cancer data sourced from the Kentucky Cancer Registry. Next, we examined colon cancer survival across various individual risk factors using Kaplan-Meier[28] survival curves. We incorporated all significant risk factors into multivariable models, employing the Cox proportional hazards model[29], the Accelerated Failure Time (AFT) model[30], the Random Survival Forest model[31], the Least Absolute Shrinkage and Selection Operator (Lasso) Regression[32], and the Elastic Net[33] procedures. Finally, we conducted subgroup analyses to further explore the risk factors that were the strongest predictors of survival outcomes. R statistical software (R Core Team, 2023) was used for all analyses[34].

The multivariate Cox proportional hazards model is a statistical method used in cancer survival studies and allows for the simultaneous analysis of multiple risk factors, estimating the risk of death while effectively controlling for confounders. This approach is particularly beneficial in identifying independent risk factors impacting survival, adjusting for the effects of other covariates. Additionally, the multivariate nature of the model enables the investigation of complex interactions between variables, providing a more comprehensive understanding of the factors affecting survival outcomes in colon cancer patients.

The Accelerated Failure Time (AFT) model is an alternative to the Cox proportional hazards model that is particularly useful for survival analysis when the proportional hazards assumption is not met

or when focusing on modeling time to an event rather than hazard rates. The AFT model assumes that covariates accelerate or decelerate survival time, which is useful in assessing how risk factors can shorten or extend patient survival time. Unlike the Cox model's focus on hazard ratios, the AFT model directly interprets time, offering clearer insights into absolute survival durations, an essential factor in understanding patient prognoses.

The Random Survival Forest (RSF) is a non-parametric, machine learning technique that extends the traditional random forest algorithm to survival data. In our study, the RSF model was utilized to predict patient outcomes based on extensive risk factors. Its major advantage lies in its ability to capture complex interactions and non-linear relationships between risk factors, dynamics that traditional models like Cox or AFT might miss. Moreover, RSF's capacity to handle censored data and missing values makes it robust for real-world clinical datasets. The model also aids in feature selection by identifying the strongest predictor variables, helping clinicians focus on key risk factors that drive survival outcomes.

Lasso and Elastic Net

The Lasso and the Elastic Net regularization techniques were utilized for the colon cancer survival data, because the number of risk factors were relatively large. Regularization is a method of constraining or penalizing model parameters to prevent overfitting, leading to better model fit and generalization.

Lasso Regression

Lasso is a regularized version of linear regression, where the model is penalized using the $L_1 norm$ (sum of the absolute values of the coefficients). The goal of using Lasso is to improve the prediction accuracy and interpretability of the model by forcing some of the coefficients that do not contribute to predicting survival to exactly zero. This essentially is equivalent to performing feature selection. The mathematical formulation of the Lasso is:

$$\widehat{\beta_{lasso}} = \underset{\beta}{argmin} ¿$$

Here, y_i is the observed target or censored time for colon cancer death in an i^{th} individual, X_i is the vector of risk factors, β_j is the coefficient of the j^{th} predictor and λ is the regularization parameter that controls the strength of penalty and *p* is the number of parameters.

Elastic Net

The Elastic Net is a regularization technique that combines both L_1 regularization (Lasso) and L_2 regularization (Ridge). It was introduced to overcome some of the limitations of Lasso, particularly when dealing with highly correlated risk factors. The Elastic Net performs better in situations where there are many predictors, some of which may be correlated. The mathematical formulation of the Elastic Net is:

$$\widehat{\beta_{elastic\,net}} = \underset{\beta}{argmin} ¿$$

Where λ_1 and λ_2 are the regularization or hyperparameters corresponding to L_1 and L_2 penalties respectively and controlling the strength of each penalty.

The Lasso procedure was chosen because we believe that only a subset of the predictors was most relevant for colon cancer survival prognosis determination. Lasso is particularly effective for sparse

data, where many features are present but only a few are significant, and its automatic feature selection enhances model interpretability. The Elastic Net was selected because the risk factors may be correlated with each other, making the Elastic Net better suited for handling multicollinearity.

RESULTS

The results from a total sample of 33,825 including 15,838 patients who were alive and censored as of December 31, 2022, are shown below. The following risk factors for colon cancer survival were investigated: age at diagnosis, treatment, sex, race, ethnicity, marital status, smoking status, cigarette pack years, tumor grade, positive lymph nodes, clinical tumor size, geographic region (Appalachian status), histology classification/type, and CS (clinical stage) lymph nodes, insurance, and seer 2018 summary classification of the tumor. Table 1(a, b) shows the distribution of the data across the risk factors while Tables 2, 3 and 4 show the comparison of the methods.

Table 1a: Distribution of colon cancer characteristics across continuous variables

	N	*Mean*	*Std*	*Median*	*Min*	*Max*
Age	33825	65.95	13.74	67.00	18.00	103.00
Cigarette Pack Years	18602	16.76	27.29	0.00	0.00	300.00
Tumor Size Clinical (mm)	7094	49.04	36.74	46.00	0.00	989.00
CS Lymph Nodes	19932	72.72	125.04	0.00	0.00	800.00

Table 1b: Distribution of colon cancer characteristics by categorical variables

Risk factors	N	Mean	Std	Median	Min	Max
Age	33825	65.95	13.74	67.00	18.00	103.00
Cigarette Pack Years	18602	16.76	27.29	0.00	0.00	300.00
Tumor Size Clinical (mm)	7094	49.04	36.74	46.00	0.00	989.00
CS Lymph Nodes	19932	72.72	125.04	0.00	0.00	800.00

Risk factors	Count	Percent
Sex	17740	53.0
Male		
Female	15730	47.0
Race	30854	92.2
White		
Black	2262	6.8
Other	354	1.1
Ethnicity	32970	98.5
Note Spanish or Latino		
Other	500	1.5
Smoking Status	11551	34.5
Non-Smoker		
Smoker	12079	36.1
Unknown/Unspecified	9840	29.4
Marital Status	4268	12.8
Married		
Single or Never Married	17905	53.5
Widowed	3969	11.9
Divorced/Separated	6139	18.3
Living with Partner or Unknown/Not reported	1142	3.6
Appalachia	23126	69.1
Non-Appalachia		
Appalachia	10344	30.9
Treatment Group	3366	10.1
No/Unknown or Refused treatment		
Surgery	23510	70.2
Chemo and or Radiation	1191	3.6
Chemo/Radiation with Surgery	2986	8.9
Other Therapies	2417	7.2

Insurance Type	289	0.9
Private Insurance		
Government-Related Programs	21563	64.4
Uninsured/Self-Pay	8293	24.8
Other Programs	3325	9.9
Histology	21841	65.3
Adenocarcinoma NOS		
Papillary Adenocarcinoma	2189	6.5
Tubulovillous Adenocarcinoma	1126	3.4
Mucinous Adenocarcinoma	2051	6.1
Other/Unspecified	6263	18.7
Positive Nodes	13975	41.8
Negative sentinel node		
Sentinel nodes Positive	8183	24.4
Unknown/Other	11312	33.8
SEER 2028 Summary Stage	4911	14.7
In situ, localized		
Regional, early spread	1645	4.9
Regional, more extensive spread	1282	3.8
Distant spread or metastasis	1563	4.7
Unknown/Unspecified	24069	71.9

Table 2: Comparison of hazard ratios across racial, diagnosis age, and geographical subgroups using Lasso regression

Risk factor	Overall	White	Black	Early	Late	Non-Appalachia	Appalachia
Age	1.035	1.035	1.033	NA	NA	1.036	1.032
Treatment (Ref = No/Unknown or Refused)							
Surgery (primary and regional)	0.309	0.304	0.321	0.297	0.28	0.327	0.289
Chemo and or radiation therapies	0.434	0.423	0.501	0.539	0.364	0.433	0.412
Surgery + (chemo/radiation)	0.264	0.26	0.286	0.269	0.202	0.277	0.246
Other therapies (immuno, endoscopic, gene, etc.)	0.622	0.61	0.68	0.653	0.43	0.68	0.558
Sex							
Female	0.865	0.859	0.897	0.929	0.853	0.855	0.901

Race (Ref = White)							
Black	1.087	NA	NA	1.176	1.027	1.069	1.127
Other	0.546	NA	NA	0.449	0.548	0.554	0.381
Ethnicity (Ref = Not Spanish or Latino)							
Other	0.96	1.000	0.866	0.796	1.000	0.978	1.000
Marital Status (Ref = Married)							
Single (never married)	0.779	0.776	0.842	0.813	0.845	0.774	0.807
Widowed	0.985	0.981	1.007	1.002	1.000	1.000	1.000
Divorced/separated	0.977	0.968	1.049	0.996	1.275	0.965	0.980
Living with partner/unknown/unreported	0.766	0.767	0.893	0.732	0.827	0.771	0.864
Smoking Status (Ref = Never smoker)							
Smoker (cigarettes, e-cigarette, cigar)	1.244	1.234	1.178	1.234	1.177	1.268	1.163
Smoker (unknown)	0.872	0.89	0.784	0.871	0.848	0.848	0.869
Cigarette Pack Years	1.002	1.002	1.000	1.004	1.001	1.001	1.003
Tumor Grade (Ref= Localized)							
Regional by direct extension	1.149	1.003	1.126	1.273	1.156	1.031	1
Regional to lymph nodes	1.545	1.338	1.219	1.955	1.504	1.407	1.218
Regional by both direct extension and regional lymph nodes	1.419	1.224	1.221	1.843	1.36	1.332	1.158
Unknown/Unstageable	1.259	1.088	1	1.289	1.244	1.076	1.179
Positive Nodes (Ref = All sentinel nodes examined are negative)							
Sentinel nodes are positive	1.538	1.555	1.668	1.838	1.502	1.559	1.586
Other/Unknown	1.917	1.909	2.132	2.724	1.749	1.938	1.883
Tumor Size	1.003	1.003	1.007	1.004	1.003	1.005	1.003
CS Lymph Nodes	1.001	1.001	1.000	1.000	1.001	1.001	1.001
Geographical Region (Ref = non-Appalachia)							
Appalachia	1.073	1.077	1.000	1.053	1.012	NA	NA
Histology (Ref = Adenocarcinoma NOS**)**							
Papillary Adenocarcinoma	0.585	0.586	0.617	0.416	0.612	0.599	0.599
Tubulovillous Adenocarcinoma	0.268	0.281	0.217	0.123	0.347	0.28	0.248
Mucinous Adenocarcinoma	1.058	1.051	1.051	1.054	1.056	1.076	1.000
Other/Unspecified	0.791	0.811	0.695	0.592	0.852	0.825	0.793
Insurance (Ref = Private Insurance)							
Government-Related Programs	0.985	1.000	1.000	0.953	1.091	1.000	0.981
Uninsured or Self-Pay	0.796	0.814	0.827	0.727	0.697	0.825	0.78
Other/Unknown Payers	1.227	1.278	1.036	1.211	1.385	1.304	1.181
SEER 2018 Summary (Ref = In situ, Localized)							
Regional, early spread	1.581	1.548	1.19	1.698	1.585	1.595	1.364
Regional, more extensive spread	1.112	1.121	0.795	1.253	1.146	1.103	1.000
Distant spread or metastasis	1.325	1.298	1.17	1.868	1.284	1.213	1.411

Unknown/Unspecified	1.974	1.949	1.621	3.297	1.797	1.976	1.832

Note: The dots (.) represent hazard ratios that were estimated as 1 (i.e., no change in hazard) by the Lasso procedure.

Survival and prevalence of colon cancer death

Figure 1 shows the annual incidence and mortality rates of colon cancer, highlighting temporal trends (slight deep in 2020 due to the COVID pandemic lockdown) in diagnosis and survival. Figure 2 illustrates the overall survival probability rates and the prevalence of colon cancer deaths over time. The overall survival rate reflects the average survival probability for patients diagnosed in a given year, whereas the prevalence of colon cancer death is defined as the ratio of total colon cancer deaths to total diagnoses in that year. It could be observed that both overall survival and the prevalence of colon cancer death agree and show the current effort being made to improve colon cancer survival. However, Figure 3 (panel A-L) shows a Kaplan-Meier curve for colon cancer survival distributions for each risk factor, highlighting statistically significant differences between the distributions of at least two subgroups for each risk factor. Figure 4 (panels A and B) shows the cross-validation results for the Lasso and Elastic Net procedures. Notably, Elastic Net results closely mirrored those of the Lasso and are omitted for brevity, with optimal lambda value determined at 0.18% for Lasso and 0.30% for Elastic Net. Figure 5 illustrates the variable importance for the risk factors in the random survival forest model. The plot highlights the order of importance, with the "red" box plots indicating significant predictors of survival. These predictors are ranked based on their contribution to the model’s ability to accurately assess survival outcomes, emphasizing the most influential variables.

Figure 1: A comparison of new CRC cases and deaths across the study period

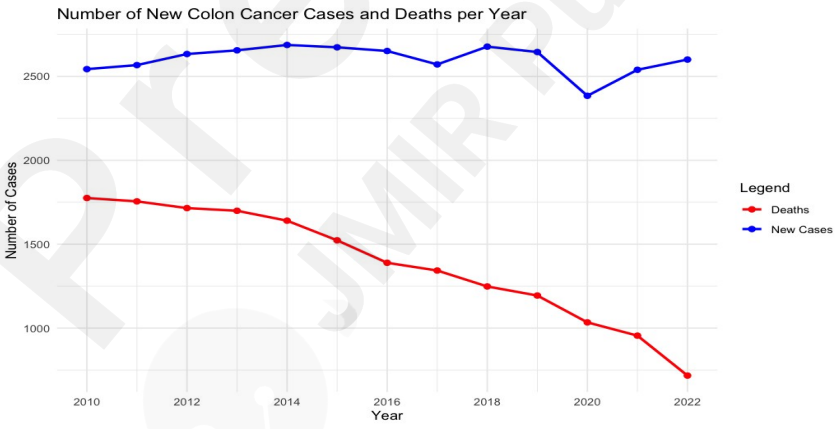


Figure 2: Understanding the CRC survival and prevalence across time.

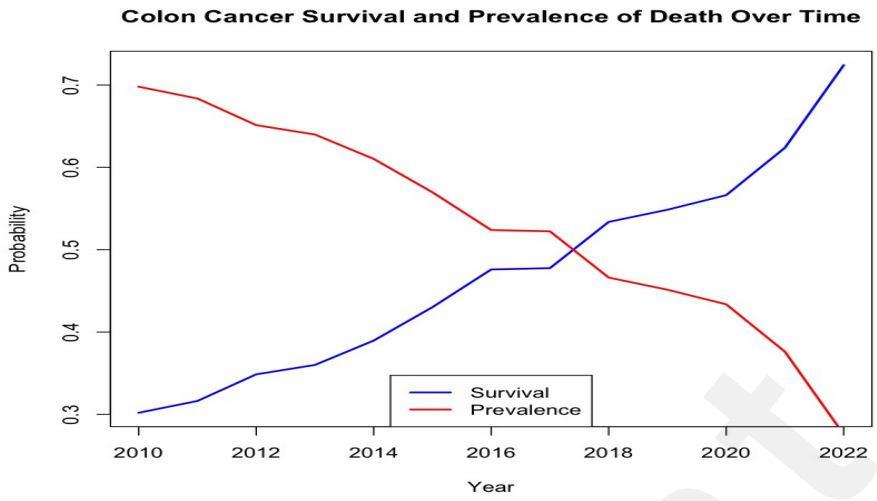
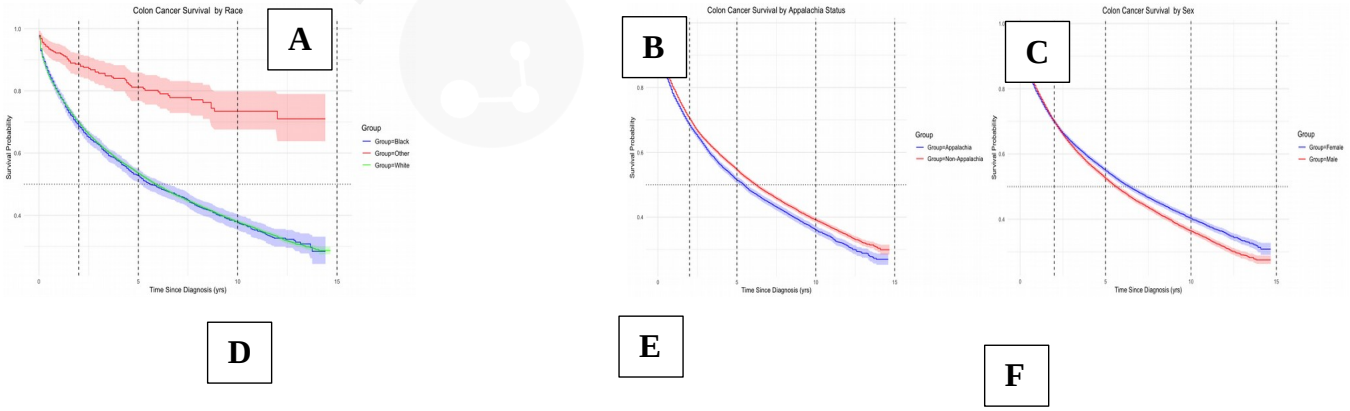


Figure 3. Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



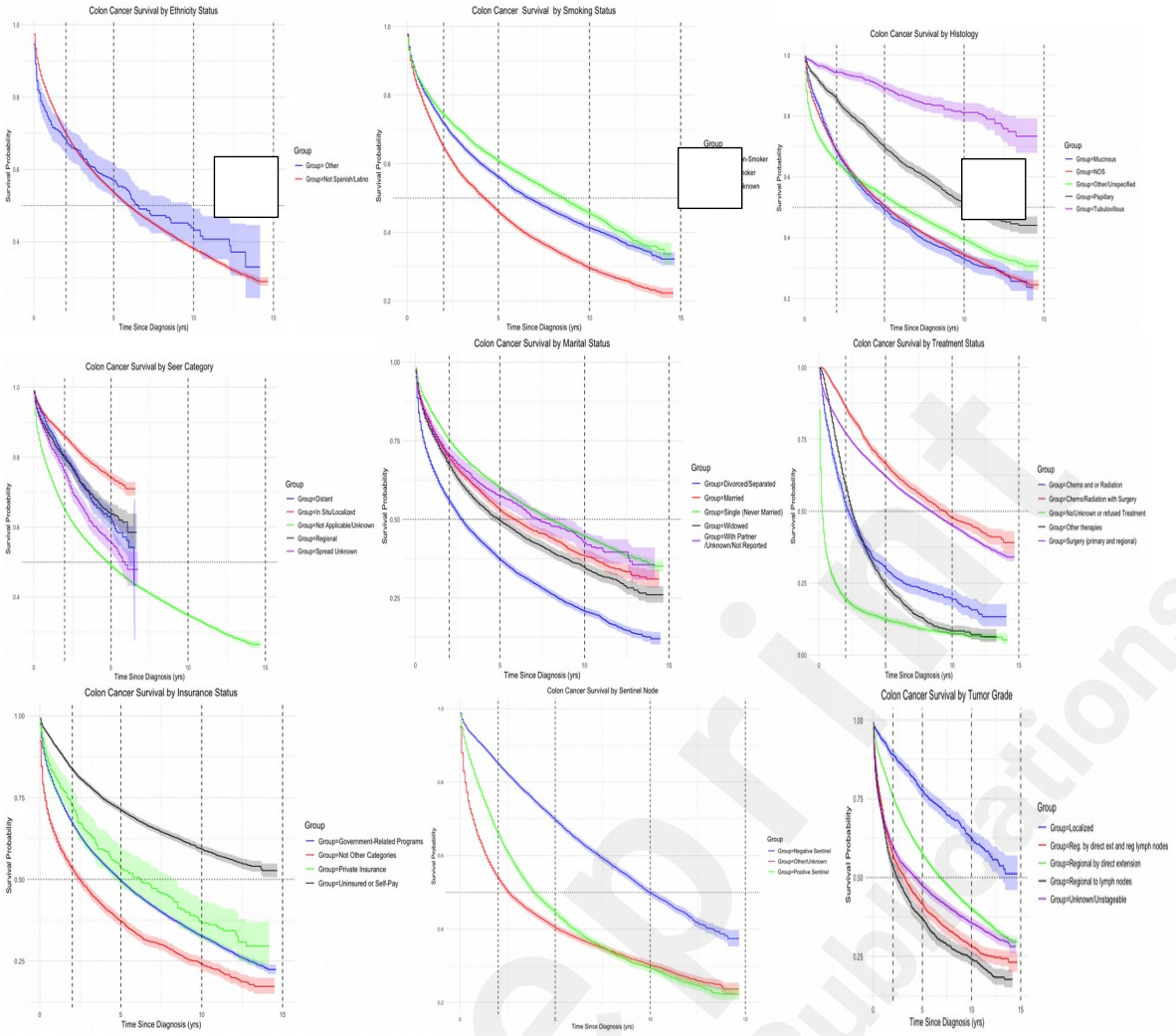


Figure 4: Cross validation for the Lasso (panel A) and the Elastic Net (panel B)

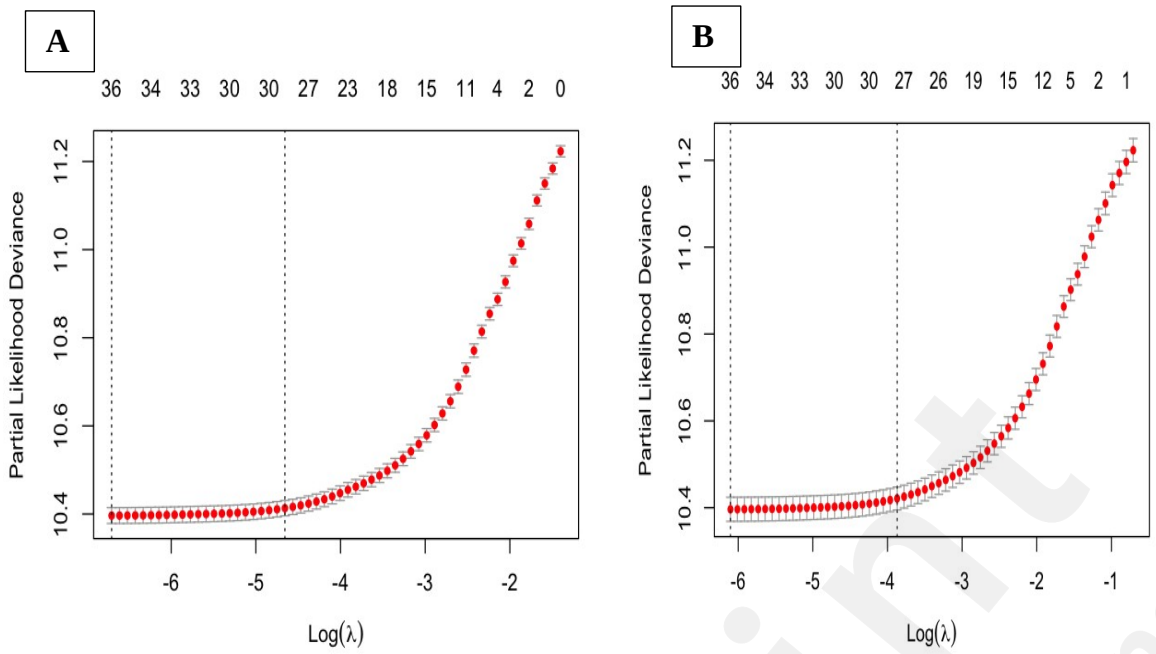
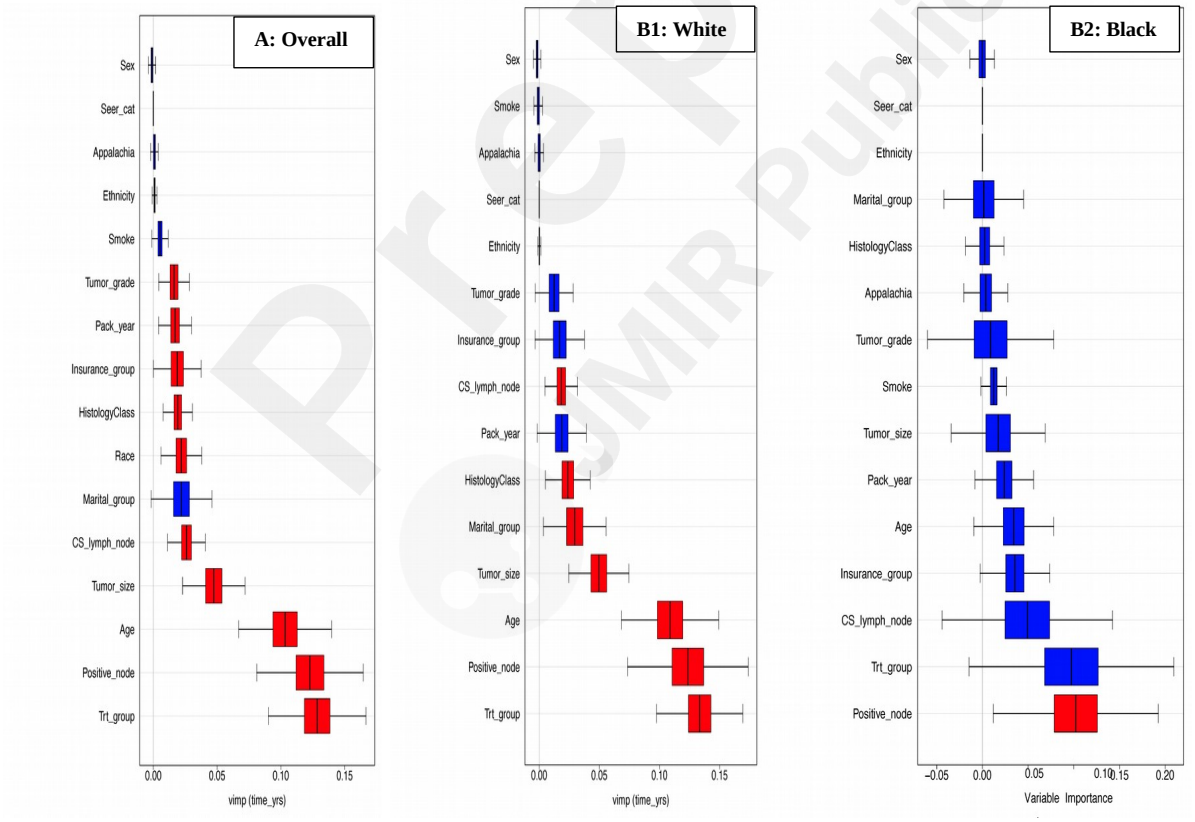
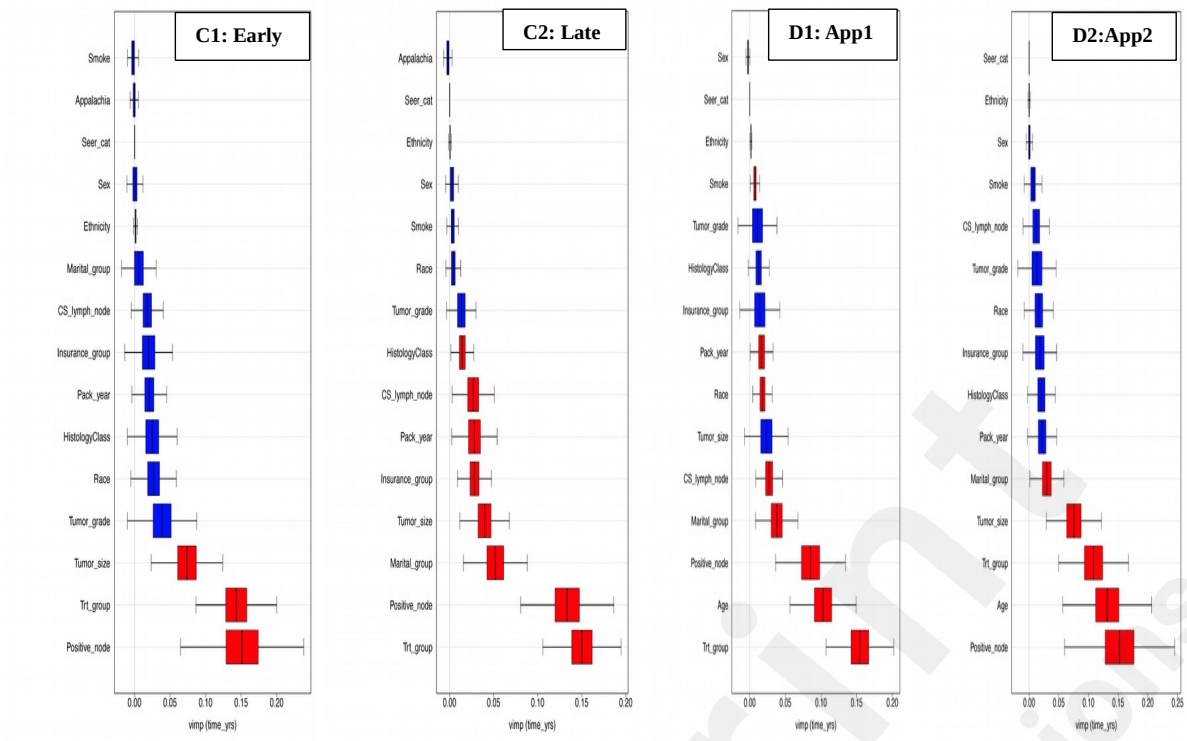


Figure 5: The overall ranking of risk factor importance and for each subgroup, as determined by the random survival forest model.





Footnote: Panel A displays the overall significant risk factors in a "red" box plot and the nonsignificant risk factors in a "blue" box plot. Panel B shows the results for the White and Black racial groups, Panel C shows that for individuals diagnosed at early or late age while panel D shows the results for living in non-Appalachia or Appalachia region. App1: Non-Appalachia, App2: Appalachia.

.

Model Results: Overall

After examining the risk factors, it was found that in the Cox model, the most significant predictors of colon cancer survival, listed in order of importance, were positive nodes, treatment group, age, tumor size, CS lymph nodes, and tumor grade, with the remaining factors having no significant impact on survival. Similarly, the AFT model identified the same set of key predictors and sex, ranked in the same order (Table 3, “Overall” column). However, in the Random Survival Forest, the importance of risk factors differed slightly, with treatment group, positive nodes, age, clinical tumor size, CS lymph nodes, marital status, race, histology, insurance, cigarette pack years, tumor grade, smoking status, ethnicity, Appalachia status, and sex following in descending order (Figure 5 panel A). The effects of ethnicity, Appalachia, and sex were relatively small compared to the other factors. The standardized out-of-bag (OOB) Continuous Ranked Probability Score (CRPS) estimate was 0.155, with a performance error of 0.254, indicating a very good fit of the model.

Table 2 presented the results for the Lasso model. For instance, the hazard ratio for overall age is 1.035, indicating that for each additional year of age, the risk of death from colon cancer increased by 3.5%. Patients who underwent surgery (primary and regional) had a 69.1 lower risk of death compared to those who refused treatment or had no treatment, representing a significant reduction in risk. The most effective treatment regimen was surgery combined with either chemotherapy or radiation therapy, which showed the greatest benefit compared to all other treatment options. Females had a 13.5 lower risk of colon cancer death compared to males (Table 2 and Figure 3 panel C). Blacks had 8.7% higher risk of colon cancer death compared to Whites, even after adjusting for other risk factors in the model (Table 2). The rest of Table 2 can be interpreted in similar ways.

Racial subgroup

The significant risk factors for predicting colon cancer survival differ across racial groups, as identified by the Cox (Table 4) and the AFT (Table 3) models. Among Whites, the most important predictors are treatment, positive nodes, age, tumor size, tumor grade, CS lymph nodes, sex, and histology. For Blacks, however, the key predictors shift slightly, with positive nodes, smoking status, insurance, age, histology, CS lymph nodes, and cigarette pack years emerging as the primary factors (Tables 3 and 4). These differences highlight the potential impact of both genetic and lifestyle factors on survival outcomes in different racial groups. In the random survival model for the Whites, the risk factors were ranked in the following order of importance: treatment group, positive nodes, age, tumor size, marital status, histology, cigarette pack years, CS lymph nodes, and tumor grade. The out-of-bag (OOB) standard concordance index (CRPS) was 0.149, with an OOB requested performance error of 0.241. The rest of the risk factors had minimal impact on survival predictions. For Blacks, the key predictors included positive nodes, treatment group, CS lymph nodes, insurance, age, cigarette pack years, tumor size, and smoking. The standard OOB CRPS was 0.184, with an OOB requested performance error of 0.327. All other risk factors contributed minimally to influence survival outcomes (Figure 5 panels B1 and B2). The above analysis was not repeated for the “Other“ racial group due to limited data.

Geographical Region

In both the AFT and Cox models, key risk factors for colon cancer survival vary between individuals living in the Appalachia region and those outside of it. For non-Appalachia residents, significant factors include treatment group, positive lymph nodes, age at diagnosis, tumor grade, marital status, and clinical tumor size (in the AFT and Cox models), with histology and CS lymph nodes also emerging as significant factor in the Cox model (Tables 3 and 4). In contrast, for Appalachia residents, survival is influenced by treatment group, age at diagnosis, positive lymph nodes, clinical tumor size, marital status, tumor grade, CS lymph node, and histology classification. The main differences between the two groups are the inclusion of CS lymph node and histology as significant factors for Appalachia residents, while clinical tumor size and smoking status are less relevant for patients in non-Appalachia region.

In the random survival forest model, the risk factors associated with colon cancer survival differ slightly between individuals living in Appalachia and those in non-Appalachia regions. For non-Appalachia residents (Figure 5 panel D1), the most influential factors, in order of significance, include treatment group, age at diagnosis, positive lymph nodes, marital status, CS lymph nodes, race, cigarette pack years and smoking status. Clinical tumor size had high impact but with high variability. Insurance, histology classification, and tumor grade also had high relevance on survival. The rest of the variables had minimal impact, with standardized OOB of 0.151 and OOB requested performance error of 0.244. In contrast, for patients in Appalachia regions, the key significant risk factors were positive lymph nodes, age at diagnosis, treatment group, clinical tumor size, and smoking status with OOB standard CRPS of 0.191 and OOB requested performance error of 0.353. Cigarette pack years, histology, insurance, race, tumor grade, CS lymph nodes, and smoking status also had high impact with high variability. The primary differences lie in the order of significance, with risk factors like treatment received and positive lymph nodes consistently being important for both non-Appalachia and Appalachian residents, but factors such as marital status, insurance, and clinical tumor size being more prominent in non-Appalachia than Appalachia.

Early vs late age diagnosis

In both the Cox and AFT models, the risk factors for colon cancer survival differ slightly between early and late stages, with some overlap and key distinctions.

For early diagnosis, the most significant risk factors across both models include positive nodes, treatment group, tumor grade, tumor size, histology and race. In the Cox model, clinical tumor size is also an important factor, while the AFT model does not highlight this as significant. For late diagnosis, both models identify treatment group, positive nodes, tumor size, marital status, and CS lymph node as key predictors. The Cox model ranked CS lymph node and tumor size higher than marital status, while the AFT model emphasizes CS lymph nodes and marital status in a similar way but with slightly different weight (Tables 3 and 4). While in the random survival forest model, the key risk factors for patients diagnosed at early age with colon cancer include positive nodes, treatment group, and tumor size. Other high-priority factors are tumor grade, race, histology, cigarette pack years, insurance, CS lymph nodes, and marital status, while the remaining factors have minimal impact on survival (Figure 5, panel C1). For patients diagnosed at an older age, the key risk factors shift slightly to include treatment group, positive nodes, marital status, clinical tumor size, insurance, cigarette pack years, CS lymph nodes, and histology. Tumor grade is also a high-priority factor. Race, smoking status, sex, ethnicity, and Appalachia status contribute minimally to survival outcomes (Figure 5, panel C2).

In summary, the risk factors for early diagnosis are largely consistent across the Cox and AFT models, with both emphasizing positive nodes, treatment group, age, tumor size, tumor grade, race, and histology. However, race is an additional significant risk factor in the AFT model that is not significant in the Cox model. For late diagnosis, both models highlight treatment group, positive nodes, age, tumor size, marital status, histology, cigarette pack years and CS lymph node though there are differences in the importance of CS lymph nodes and marital status. In the random survival forest model, many of the same risk factors overlap, but there are notable distinctions. Clinical tumor grade, race, histology, cigarette pack years, insurance status, CS lymph node, and marital status had more variability in the early diagnosis group, while clinical tumor size, marital status, tumor size, insurance status, cigarette pack years, emerge as key additional factors for patients diagnosed at a later age beyond treatment and positive node.

DISCUSSION

The primary risk factors for colon cancer death included positive lymph nodes, age, treatment received, clinical tumor size, tumor grade, positive sentinel lymph node, smoking, geographic region, and marital status across all the statistical models. The highest risk of death was observed among individuals who received no treatment or refused treatment, compared to those who underwent any other form of treatment. Social determinants of health such as insurance (which serves as a proxy for poverty), Appalachia region, marital status, and modifiable risk factors such as smoking significantly impact survival. Not having insurance, smoking, living in Appalachia region, and being married are associated with higher risk of dying, highlighting the importance of these factors in survival outcomes. Consistent with prior studies, these results affirm that men have a higher risk of colon cancer death than women[35-37].

Overall, smokers had a 24 higher risk of mortality compared to non-smokers: a finding consistent with the results reported by Tsao et al[38]. Additionally, Appalachian residents had a 7% higher mortality risk than non-Appalachians (Table 2). In non-Appalachian residents, the hazard of death was about 21% higher for those with distant metastasis compared to those with in situ localized cancer diagnosis. In contrast, Appalachian residents experienced an even greater 20% increase, resulting in a 41 higher risk of death, indicating a more severe association between geographic factors and cancer spread, like those reported by Wang et al[39].

When comparing significant risk factors across racial groups in the Cox model, it was evident that while the type of treatment received was a significant predictor of survival for the White group, it did not appear to be a useful predictor for the Black group (Table 4). This disparity may highlight the possibility that Black patients often seek treatment later in the disease course, potentially reducing the effectiveness of the treatment they receive. However, modifiable risk factors such as smoking, cigarette pack years, and insurance status were found to impact survival. This suggests that addressing these factors could help narrow the gap in colon cancer survival between White and Black patients.

In the Random Survival Forest model, differences emerged in how treatment and socioeconomic factors (like insurance and smoking history) influence survival outcomes across racial groups. Both groups shared predictors like positive nodes and clinical tumor size, tumor grade, but Whites emphasized treatment received and positive note, while the Black's survival was more impacted by positive node and treatment received. Other key effects include smoking and insurance influenced survival compared to whites. This contrast suggests that while biological factors (such as tumor characteristics) are universally relevant, socioeconomic and access disparities exacerbate survival gaps between racial groups. Overall, positive lymph nodes, treatment, age at diagnosis, CS lymph nodes, tumor size, tumor grade, and histology were important risk factors in both Appalachian and non-Appalachian areas of Kentucky. CS lymph nodes, however, was more influential within the non-Appalachian population compared to patients in Appalachia.

For the Cox model, while positive nodes, treatment group, and histology were consistently significant for both early and late diagnosis, tumor characteristics and social risk factors like marital status vary between two the groups (early vs late diagnosis). These findings suggest that we need to integrate critical risk factors into individualized prognostic models. Unfortunately, limitations in level II data precluded analysis of such modifiable risks as alcohol consumption, obesity, familial associations, inflammatory bowel diseases, and consumption of processed and red meats, which are known to increase colorectal cancer risk.

CONCLUSION

This study evaluated multiple survival analysis models, including the Cox Proportional Hazards Model (Cox model), the Accelerated Failure Time (AFT) model, Random Survival Forests (RSF), the Lasso regression, and Elastic Net. Each of these models has distinct strengths and assumptions, making them suitable for different aspects of our research questions. Both Lasso and Elastic Net are powerful regularization techniques that helped improve the model generalization, interpretability, and predictive accuracy.

Colon cancer survival and prognosis

Different subgroups may have varying priorities when it comes to the risk factors that influence colon cancer survival and prognosis determination, rather than relying on a one-size-fits-all

approach. Understanding these subgroup-specific risk factors enhances our ability to tailor initiatives, interventions, and preventive measures for different populations, particularly in terms of screening, treatment, diagnosis, and prognosis.

By using the results from the Cox method, clinicians and researchers can better understand which risk factors are important in predicting a patients' risks and guide personalized treatment strategies. Using the AFT model for colon cancer survival analysis provides valuable insights, especially when the goal is to predict how long patients are likely to survive based on their individual characteristics, which can inform personalized treatment strategies and improve a patients' risk management technique. By leveraging the predictive power of RSF, the Lasso and Elastic Net techniques, researchers and clinicians can develop more accurate and individualized survival risk predictions for colon cancer patients, ultimately supporting personalized treatment plans and improving decision-making in clinical practice. For example, in the AFT model, among individuals diagnosed at an early age, smokers experience death 10% sooner than never-smokers. In contrast, among those diagnosed at a late age, smokers experience death 13% sooner than never-smokers. This suggests that the modifiable risk factor of smoking might be targeted more urgently among those diagnosed with CRC at older age than among those diagnosed at an older age.

Colon cancer remains a significant health challenge across many subpopulations, with survival outcomes strongly influenced by both early detection and the management of various risk factors. Modifiable lifestyle factors, such as smoking, lack of insurance, and geographic location (which, although not directly modifiable by individuals can be addressed through public health interventions), are crucial in mitigating the risk of colorectal cancer. Additionally, non-modifiable factors, including age, genetic predispositions, and medical conditions such as inflammatory bowel disease, further shape an individual's risk profile. Early detection through screening remains the most effective strategy for improving survival[40,41], as it significantly increases the likelihood of successful treatment when cancer is diagnosed at an early stage. By addressing these risk factors for different subgroups through improved preventive strategy and intervention and personalized treatment approaches, we can significantly reduce the global burden of colon cancer.

Acknowledgments
We thank Jaclyn McDowell, DrPH at the Kentucky Cancer Registry for her assistance in extracting the colon cancer data and Jerod Stapleton, PhD for reading through this manuscript.

Conflict of Interest
None declared.

References

1. Xi Y, & Xu P, (2021). Global colorectal cancer burden in 2020 and projections to 2040. Translational oncology; 14(10): 101174. doi:10.1016/j.tranon.2021.101174.
2. Gandomani HS, Aghajani M, Mohammadian-Hafshejani A, Tarazoj AA, Pouyesh V, & Salehiniya H, (2017). Colorectal cancer in the world: incidence, mortality and risk factors. Biomedical Research and Therapy; 4(10): 1656-1675.
3. Centers for Disease Control and Prevention (CDC). Colorectal cancer statistics. Available at: https://www.cdc.gov/cancer/colorectal/statistics/index.htm.
4. Kentucky Cancer Consortium. Kentucky cancer burden report. Available at:

https://kycancercouncil.org.
5. American Cancer Society. Cancer facts & figures 2020. American Cancer Society; 2020. Available at: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html.
6. Kenamond MC, Mourad WF, Randall ME, & Kaushal, A. (2022). No oncology patient left behind: Challenges and solutions in rural radiation oncology. The Lancet Regional Health–Americas, 13.
7. Paskett ED, Fisher JL, Lengerich EJ, Schoenberg NE, Kennedy SK, Conn ME, Roberto KA, Dwyer SK, Fickle D, Dignan M, Disparities in Underserved White Populations: The Case of Cancer-Related Disparities in Appalachia, The Oncologist, Volume 16, Issue 8, August 2011, Pages; 1072–1081, https://doi.org/10.1634/theoncologist.2011-0145
8. Wolbert T, Leigh EC, Barry R, Thompson EC, Gress T, Ajmera A, & Arrington, A. K. (2018). Later stage disease and earlier onset of rectal cancer: epidemiology and outcomes comparison of rectal cancer in a rural Appalachian area to state and national rates. The American Surgeon, 84(7), 1229-1235.
9. Lengerich EJ, Tucker TC, Powell RK, Colsher P, Lehman E, Ward AJ, & Wyatt, S. W. (2005). Cancer incidence in Kentucky, Pennsylvania, and west virginial: Disparities in Appalachia. The Journal of Rural Health, 21(1), 39-47. doi: /10.1111/j.1748-0361.2005.tb00060.x
10. Cancer Research UK. Colorectal cancer survival statistics. Available at: https://www.cancerresearchuk.org. 2020.
11. Siegel RL, Miller KD, & Jemal A, (2018). Cancer statistics, 2018. CA: a cancer journal for clinicians; 68(1): 7-30. doi: 10.3322/caac.21442.
12. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med. (2015); (372): 1601-1613. doi: 10.1056/NEJMra1401510.
13. Giardiello FM, Allen JI, Axilbund JE, Boland CR, Burke CA, Burt RW, Church JM, Dominitz JA, Johnson DA, Kaltenbach T, Levin TR, Lieberman DA, Robertson DJ, Syngal S, Rex DK. Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-society Task Force on colorectal cancer. Am J Gastroenterol. 2014 Aug;109(8):1159-79. doi: 10.1038/ajg.2014.186. Epub 2014 Jul 22. PMID: 25070057.
14. Boland CR, Goel A. Somatic mutations and epigenetic changes in the pathogenesis of colorectal cancer. Annu Rev Pathol. (2010); (5): 461-480. doi: 10.1146/annurev-pathol-012809-153533.
15. Baena, R., & Salinas, P. (2015). Diet and colorectal cancer. Maturitas; 80(3): 258-264. doi: 10.1016/j.maturitas.2014.12.017.
16. International Agency for Research on Cancer. IARC monographs on the evaluation of carcinogenic risks to humans: Volume 114, red meat and processed meat. Available at: https://monographs.iarc.who.int. 2015.
17. de Menezes RF, Bergmann A, & Thuler LCS, (2013, September 30). Alcohol Consumption and Risk of Cancer: A Systematic Literature Review. Asian Pacific Journal of Cancer Prevention; Asian Pacific Organization for Cancer Prevention. doi: 10.7314/apjcp.2013.14.9.4965.
18. Martínez ME, McPherson RS, Annegers JF, & Levin B, (1995). Cigarette smoking and alcohol consumption as risk factors for colorectal adenomatous polyps. JNCI: Journal of the National Cancer Institute; 87(4): 274-279. doi: 10.1093/jnci/87.4.274.
19. Friedenreich CM, Shaw E, Neilson HK, & Brenner DR, (2017). Epidemiology and biology of physical activity and cancer recurrence. Journal of Molecular Medicine; (95): 1029-1041. doi: 10.1007/s00109-017-1558-9
20. Wiseman M. The Second World Cancer Research Fund/American Institute for Cancer

Research Expert Report. Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective: Nutrition Society and BAPEN Medical Symposium on 'Nutrition support in cancer therapy.' *Proceedings of the Nutrition Society*. 2008;67(3):253-256. doi:10.1017/S002966510800712X

21. Ma Y, Yang Y, Wang F, Zhang P, Shi C, Zou Y, & Qin H, (2013). Obesity and risk of colorectal cancer: a systematic review of prospective studies. PloS one; 8(1): e53916. doi: 10.1371/journal.pone.0053916.
22. Giovannucci E, (2003). Nutrition, insulin, insulin-like growth factors and cancer. Hormone and Metabolic Research, 35(11/12): 694-704. doi: 10.1055/s-2004-814147.
23. Gibson TM, Park Y, Robien K, Shiels MS, Black A, Sampson JN, ... & Morton LM, (2014). Body mass index and risk of second obesity-associated cancers after colorectal cancer: a pooled analysis of prospective cohort studies. Journal of clinical oncology; 32(35): 4004-4011. doi: 10.1200/JCO.2014.56.8444
24. Kahi CJ, Imperiale TF, Juliar BE, & Rex DK, (2009). Effect of screening colonoscopy on colorectal cancer incidence and mortality. Clinical gastroenterology and hepatology; 7(7): 770-775. doi: 10.1016/j.cgh.2008.12.030.
25. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, ... & Jemal A, (2020). Colorectal cancer statistics, 2020. CA: a cancer journal for clinicians; 70(3): 145-164. doi: 10.3322/caac.21601
26. Albadari N, Xie Y, & Li W, (2024). Deciphering treatment resistance in metastatic colorectal cancer: roles of drug transports, EGFR mutations, and HGF/c-MET signaling. Frontiers in Pharmacology; (14): 1340401. doi: 10.3390/ijms24021344.
27. National Cancer Institute. SEER Summary Stage Manual 2018. Available at: https://seer.cancer.gov/tools/ssm/2018-Summary-Stage-Manual.pdf. Accessed [February 13, 2025].
28. Kaplan EL, & Meier P, (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282): 457-481.
29. Cox DR, (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2): 187-202. https://www.jstor.org/stable/2985181
30. Wei LJ (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in medicine, 11(14-15): 1871-1879. doi: 10.1002/sim.4780111409
31. Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. (2008): 841-860. doi: 10.1214/08-AOAS169
32. Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1): 267-288. doi:10.1111/j.2517-6161.1996.tb02080.x
33. Zou H & Hastie T, (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2): 301-320. doi:10.1111/j.1467-9868.2005.00503.x
34. Team, R. C. (2020). RA language and environment for statistical computing, R Foundation for Statistical. Computing.
35. Singh PN, Fraser GE (1998) Dietary risk factors for colon cancer in a low-risk population. Am J Epidemiol; (148):761–774. doi: 10.1093/oxfordjournals.aje.a009697.
36. Giovannucci E (2002) Obesity, gender, and colon cancer; Gut (51):147. doi: 10.1136/gut.51.2.147
37. Ning Y, Wang L, Giovannucci EL, (2010). A quantitative analysis of body mass index and colorectal cancer: findings from 56 observational studies; Obes Rev (11):19–30. doi: 10.1111/j.1467-789X.2009.00613.x.
38. Tsoi KK, Pau YY, Wu WK, Chan FKL, Griffiths SM, Sung JJ (2009) Cigarette smoking and

the risk of colorectal cancer: a meta-analysis of prospective cohort studies. Clin Gastroenterol Hepatol; (7):682–688. doi: 10.1016/j.cgh.2009.02.016.
39. Wang CC, Farmer T, Garland-Kledzik M, Magge DR. Disparities in Advanced Stage Colorectal Cancer Outcomes in Appalachia: A Comprehensive Review. The American Surgeon™. 2025; 0(0). doi:10.1177/00031348241312124.
40. NCI (2024) Colorectal cancer screening (PDQ®). National Cancer Institute, Rockville, MD. https://www.cancer.gov/types/colorectal/hp/colorectal-screening-pdq
41. NCI (2024) Colorectal cancer prevention (PDQ®). National Cancer Institute, Rockville, MD. http://www.cancer.gov/cancertopics/pdq/prevention/colorectal/HealthProfessional/page2#Section_995.

Supplementary

Table 3: Log survival time estimates for the overall population and across the different subgroups.

	Overall		White		Black		Diagnosis				Appalachia			
							Early		Late		No		Yes	
Risk factor	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**
Intercept	3.287 (0.8983)	2.53E-04	3.172 (0.9129)	5.12E-04	5.413 (2.2062)	0.01411	1.378 (1.2408)	2.67E-01	1.821 (1.2081)	1.32E-01	3.094 (1.2235)	1.15E-02	4.514 (1.393)	1.19E-03
Age	-0.026 (0.0047)	2.14E-08	-0.026 (0.0049)	1.83E-07	-0.069 (0.0233)	0.0033	NA	NA	NA	NA	-0.023 (0.0053)	1.42E-05	-0.050 (0.0104)	1.57E-06
Treatment (Ref = No/Unknown or Refused)														
Surgery (primary and regional)	1.540 (0.1901)	5.51E-16	1.613 (0.2065)	5.75E-15	-0.146 (0.5923)	0.80533	1.560 (0.3981)	8.89E-05	1.637 (0.2316)	1.58E-12	1.556 (0.2254)	5.15E-12	1.858 (0.384)	1.31E-06
Chemo and or radiation therapies	1.809 (0.2167)	6.77E-17	1.934 (0.2256)	9.97E-18	0.377 (0.7359)	0.60844	1.507 (0.4402)	6.19E-04	2.088 (0.2708)	1.25E-14	1.917 (0.2703)	1.30E-12	2.157 (0.4033)	8.93E-08
Surgery + (chemo/ radiation)	1.617 (0.2217)	3.10E-13	1.731 (0.2352)	1.88E-13	-0.424 (0.723)	0.55577	1.827 (0.4437)	3.84E-05	1.737 (0.2636)	4.49E-11	1.600 (0.2708)	3.43E-09	2.065 (0.411)	5.06E-07
Other therapies (immuno, endoscopic, gene, etc.)	0.574 (0.2058)	5.31E-03	0.669 (0.2211)	2.49E-03	-0.451 (0.806)	0.57544	0.706 (0.3986)	7.65E-02	0.763 (0.2585)	3.17E-03	0.331 (0.2391)	1.66E-01	1.574 (0.4171)	1.61E-04
Sex														
Female	0.210	3.88	0.187	7.81	0.754	0.1	0.343	6.84	0.120	3.47	0.225	5.51	-0.21	3.10

	(0.1016)	E-02	(0.1058)	E-02	(0.4653)	052	(0.1881)	E-02	(0.1274)	E-01	(0.1172)	E-02	13 (0.2083)	E-01
Race (Ref = White)														
Black	0.134 (0.181)	4.60E-01	NA	NA	NA	NA	0.145 (0.3156)	6.46E-01	0.077 (0.2345)	7.42E-01	0.166 (0.1906)	3.83E-01	0.011 (0.5992)	9.85E-01
Other	-0.305 (0.4872)	5.27E-01	NA	NA	NA	NA	-1.233 (0.6254)	4.86E-02	0.278 (0.7852)	7.24E-01	-0.184 (0.4815)	7.02E-01	NA	NA
Ethnicity (Ref = Not Spanish or Latino)														
Other	0.252 (0.6189)	6.84E-01	0.237 (0.6175)	7.01E-01	NA	NA	1.003 (1.0564)	3.43E-01	-0.133 (0.7792)	8.64E-01	0.575 (0.7559)	4.47E-01	-0.066 (1.0092)	9.48E-01
Marital Status (Ref = Married)														
Single (never married)	0.224 (0.1431)	1.17E-01	0.207 (0.1541)	1.80E-01	-0.711 (0.4865)	0.144	0.225 (0.212)	2.88E-01	0.079 (0.208)	7.05E-01	0.149 (0.1618)	3.59E-01	0.883 (0.3012)	3.37E-03
Widowed	-0.086 (0.1772)	6.26E-01	-0.140 (0.1869)	4.52E-01	0.281 (0.7848)	0.7207	-0.163 (0.2623)	5.35E-01	-0.209 (0.2526)	4.08E-01	0.004 (0.2102)	9.87E-01	0.200 (0.3321)	5.46E-01
Divorced/ separated	-0.249 (0.1702)	1.44E-01	-0.311 (0.1828)	8.94E-02	0.446 (0.598)	0.4562	-0.359 (0.4708)	4.46E-01	-0.497 (0.2153)	2.11E-02	-0.426 (0.1934)	2.75E-02	0.669 (0.3645)	6.66E-02
Living with partner/unknown/unreported	-0.495 (0.3605)	1.70E-01	-0.533 (0.3826)	1.64E-01	-0.794 (0.9905)	0.4226	-0.687 (0.7633)	3.68E-01	-0.575 (0.4404)	1.92E-01	-0.718 (0.4241)	9.07E-02	0.510 (0.69)	4.60E-01
Smoking Status (Ref = Never smoker)														
Smoker (cigarettes, e-cigarette, cigar)	-0.121 (0.1266)	3.40E-01	-0.152 (0.1321)	2.51E-01	1.366 (0.4502)	0.0024	-0.099 (0.2191)	6.52E-01	-0.134 (0.1601)	4.02E-01	-0.011 (0.1459)	9.43E-01	-0.503 (0.2587)	5.19E-02
Smoker (unknown)	-1.796 (1.0746)	9.46E-02	NA	NA	-1.638 (0.9538)	0.0859	NA	NA	-1.618 (1.1183)	1.48E-01	-1.835 (1.0687)	8.61E-02	NA	NA
Cigarette Pack Years	-0.001 (0.0021)	5.70E-01	-7e-04 (0.0022)	7.42E-01	-0.019 (0.0111)	0.0866	-0.004 (0.0044)	3.68E-01	1E-04 (0.0026)	9.63E-01	-0.002 (0.0025)	5.43E-01	-4E-04 (0.0043)	9.24E-01
Tumor Grade (Ref= Localized)														
Regional by direct	-	2.	-	5.	-	0.	-	1.	-	1.	-	7.	-	3.

extension	0.657 (0.2894)	32E-02	0.585 (0.3053)	56E-02	0.143 (0.8674)	0.8695	0.690 (0.4842)	54E-01	0.538 (0.3613)	37E-01	0.657 (0.3642)	12E-02	1.026 (0.4901)	64E-02
Regional to lymph nodes	-1.031 (0.3201)	1.28E-03	-1.011 (0.3359)	2.61E-03	0.238 (1.2723)	0.8517	-1.479 (0.5522)	7.81E-03	-0.848 (0.3981)	3.31E-02	-1.179 (0.4011)	3.28E-03	-1.148 (0.5286)	2.99E-02
Regional by both direct extension and regional lymph nodes	-0.853 (0.3184)	7.39E-03	-0.810 (0.3367)	1.61E-02	-1.046 (0.9808)	0.2863	-1.704 (0.5635)	2.49E-03	-0.505 (0.3923)	1.98E-01	-0.879 (0.3998)	2.79E-02	-1.084 (0.5347)	4.26E-02
Unknown/ Unstageable	-0.482 (0.3054)	1.15E-01	-0.330 (0.3275)	3.14E-01	-0.267 (0.9357)	0.7752	-1.137 (0.5212)	2.91E-02	-0.184 (0.3879)	6.35E-01	-0.354 (0.389)	3.63E-01	-0.983 (0.5039)	5.11E-02
Positive Nodes (Ref = All sentinel nodes examined are negative)														
Sentinel nodes are positive	-0.298 (0.1586)	5.99E-02	-0.248 (0.1668)	1.37E-01	-0.346 (0.5013)	0.4901	-0.290 (0.2746)	2.91E-01	-0.202 (0.2067)	3.29E-01	-0.243 (0.1956)	2.15E-01	-0.232 (0.2752)	3.98E-01
Other/Unknown	-1.289 (0.1434)	2.51E-19	-1.316 (0.1492)	1.16E-18	-2.011 (0.5881)	0.0006	-1.379 (0.2587)	9.77E-08	-1.281 (0.181)	1.49E-12	-1.178 (0.1692)	3.25E-12	-1.517 (0.2752)	3.56E-08
Clinical Tumor Size	-0.003 (7e-04)	1.86E-06	-0.003 (7e-04)	7.85E-06	-0.001 (0.0067)	0.9447	-0.003 (9E-04)	1.94E-03	-0.005 (0.0019)	5.52E-03	0.005 (0.0021)	2.53E-02	-0.003 (8E-04)	1.15E-03
Geographical Region (Ref = non-Appalachia)														
Appalachia	0.050 (0.1101)	6.49E-01	0.047 (0.1121)	6.76E-01	-0.290 (0.5181)	0.576	0.164 (0.1971)	4.05E-01	0.095 (0.1358)	4.86E-01	NA	NA	NA	NA
Histology (Ref = Adenocarcinoma NOS**)**														
Papillary Adenocarcinoma	0.102 (0.2147)	6.34E-01	0.037 (0.2184)	8.67E-01	0.605 (0.9711)	0.5331	0.585 (0.4108)	1.54E-01	-0.111 (0.2588)	6.69E-01	-0.012 (0.2486)	9.61E-01	-9E-04 (0.416)	9.98E-01
Tubulovillous Adenocarcinoma	0.676 (0.4473)	1.31E-01	0.848 (0.4796)	7.71E-02	1.217 (1.298)	0.3484	1.660 (0.7681)	3.07E-02	0.625 (0.5585)	2.63E-01	0.101 (0.5395)	8.51E-01	1.870 (0.7959)	1.88E-02
Mucinous Adenocarcinoma	-0.342 (0.1775)	5.42E-02	-0.363 (0.1801)	4.40E-02	-0.785 (0.9421)	0.4046	0.293 (0.3596)	4.15E-01	-0.433 (0.2128)	4.18E-02	-0.466 (0.1916)	1.50E-02	0.271 (0.4639)	5.60E-01
Other/Unspecified	0.006 (0.1	9.67E-	-0.045	7.51E-	0.975 (0.4	0.002	0.418 (0.2	1.33E-	-0.139	4.05E-	-0.040	8.16E-	-0.042	8.50E-

	378)	01	(0.1419)	01	436)	8	781)	01	(0.1666)	01	(0.1719)	01	(0.2238)	01
CS Lymph Nodes	-0.001 (5e-04)	2.47E-03	-0.001 (5e-04)	8.18E-03	-0.004 (0.014)	0.008	-3E-04 (7E-04)	6.88E-01	-0.002 (7E-04)	7.78E-04	-0.001 (6E-04)	5.01E-02	-0.002 (7E-04)	1.35E-02
Insurance (Ref = Private Insurance)														
Government-Related Programs	0.393 (0.7539)	6.03E-01	0.340 (0.7524)	6.52E-01	1.870 (0.6483)	0.0039	0.749 (1.0321)	4.68E-01	0.010 (1.1026)	9.93E-01	0.441 (1.0636)	6.78E-01	0.398 (1.0225)	6.97E-01
Uninsured or Self-Pay	0.910 (0.7581)	2.30E-01	0.885 (0.7570)	2.43E-01	1.790 (0.6881)	0.0093	1.115 (1.0277)	2.78E-01	0.941 (1.1216)	4.01E-01	0.953 (1.0606)	3.69E-01	0.814 (1.0507)	4.39E-01
Other/Unknown Payers	0.051 (0.7686)	9.47E-01	0.082 (0.7695)	9.15E-01		NA	0.475 (1.0817)	6.60E-01	-0.318 (1.1142)	7.75E-01	-0.171 (1.0796)	8.74E-01	0.454 (1.0561)	6.68E-01
Log(scale)	0.048 (0.0375)	2.01E-01	0.044 (0.0392)	2.60E-01	-0.292 (0.1341)	0.0297	-0.012 (0.0687)	8.57E-01	0.074 (0.0445)	9.71E-02	0.037 (0.0437)	3.96E-01	-0.034 (0.0722)	6.41E-01

Footnote: Est (std): Estimated log survival time from the accelerated failure time model (standard deviation), P-val: Adjusted p-value. Notice that the negative sign means that as the covariate increases, the probability of experiencing death increases (survival time is shorter).

Table 4: Log hazard ratio estimates for the overall population and across the subgroups.

	Overall		**White**		**Black**		**Early Diagnosis**		**Late Diagnosis**		**Non-Appalachia**		**Appalachia**	
Risk factors	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**	**Est (Std)**	**P-Val**
Age	0.026 (0.0045)	1.2268E-08	0.025 (0.0047)	1.1807E-07	0.088 (0.0316)	0.005 26566	NA	NA	NA	NA	0.023 (0.0052)	1.1525E-05	0.053 (0.0107)	8.8763E-07
Treatment (Ref = No/Unknown or Refused)														
Surgery (primary and regional)	-1.476 (0.1839)	1.0384E-15	-1.553 (0.2016)	1.3559E-14	0.010 (0.8018)	0.98993533	-1.640 (0.405)	5.1105E-05	-1.526 (0.2196)	3.6374E-12	-1.502 (0.2207)	1.0095E-11	-1.965 (0.4145)	2.1392E-06
Chemo and or radiation therapies	-1.748 (0.21	1.2293E-16	-1.875 (0.22	3.0569E-17	-0.782 (0.99	0.43178175	-1.603 (0.45	0.00041565	-1.962 (0.25	2.912E-14	-1.857 (0.2661)	2.9558E-12	-2.295 (0.4405)	1.8897E-07

	11)		2)		53)		43)		81)					
Surgery + (chemo/radiation)	-1.541 (0.2132)	4.9807E-13	-1.659 (0.2282)	3.6312E-13	0.395 (0.9838)	0.68776855	-1.925 (0.4497)	1.8734E-05	-1.598 (0.2482)	1.1951E-10	-1.537 (0.2634)	5.3718E-09	-2.183 (0.444)	8.775E-07
Other therapies (immuno, endoscopic, gene, etc.)	-0.537 (0.1979)	0.00660985	-0.63 (0.2146)	0.00330889	0.560 (1.0734)	0.60209479	-0.795 (0.4102)	0.05244838	-0.676 (0.2419)	0.00518855	-0.305 (0.2317)	0.188536	-1.658 (0.4496)	0.00022735
Sex														
Female	-0.201 (0.0968)	0.03767692	-0.180 (0.1013)	0.07553708	-0.861 (0.627)	0.16964131	-0.341 (0.1896)	0.07220556	-0.112 (0.1184)	0.34440 17	-0.214 (0.1129)	0.05757586	0.223 (0.2163)	0.30270 06
Race (Ref = White)														
Black	-0.127 (0.1728)	0.46112 41	NA	NA	NA	NA	-0.177 (0.3215)	0.58280 42	-0.063 (0.2182)	0.77195 91	-0.161 (0.1842)	0.38121 96	-0.008 (0.6294)	0.99046 09
Other	0.304 (0.4602)	0.50897 18	NA	NA	NA	NA	1.211 (0.6331)	0.05566379	-0.225 (0.7296)	0.75745 27	0.19 (0.4641)	0.68153 12	NA	NA
Ethnicity (Ref = Not Spanish or Latino)														
Other	-0.236 (0.5899)	0.68893 08	-0.226 (0.5909)	0.70176 64	NA	NA	-0.935 (1.0675)	0.38104 44	0.165 (0.723)	0.81997 19	-0.559 (0.728)	0.44284 63	0.126 (1.0459)	0.90429 58
Marital Status (Ref = Married)														
Single (never married)	-0.218 (0.1366)	0.11030 72	-0.201 (0.1477)	0.17399 45	0.878 (0.6634)	0.18566773	-0.221 (0.2154)	0.30424 59	-0.09 (0.1932)	0.64245 57	-0.150 (0.1562)	0.33668 59	-0.931 (0.3134)	0.00297472
Widowed	0.089 (0.169)	0.59830 71	0.145 (0.179)	0.41761 02	-0.450 (1.026)	0.66065474	0.182 (0.2651)	0.49304 78	0.189 (0.2346)	0.41926 43	-0.001 (0.2027)	0.99462 25	-0.193 (0.3445)	0.57586 61
Divorced/ separated	0.233 (0.1622)	0.15177 88	0.297 (0.1749)	0.08896344	-0.840 (0.8373)	0.31583105	0.297 (0.4795)	0.53493 86	0.453 (0.1993)	0.02313812	0.402 (0.186)	0.03085635	-0.698 (0.3768)	0.06399579
Living with partner/unknown/unreported	0.481 (0.344)	0.16186 84	0.524 (0.3665)	0.15316 52	1.067 (1.3668)	0.43502278	0.681 (0.7692)	0.37608 25	0.526 (0.4095)	0.19903 97	0.703 (0.4092)	0.08562187	-0.593 (0.7168)	0.40775 64

Smoking Status (Ref = Never smoker)														
Smoker (cigarettes, e-cigarette, cigar)	0.113 (0.1208)	0.35070 89	0.144 (0.1266)	0.25678 69	-1.788 (0.6389)	0.00513 185	0.089 (0.2228)	0.68820 67	0.123 (0.1488)	0.40959 23	0.008 (0.1406)	0.95669 05	0.510 (0.2678)	0.05672 918
Smoker (unknown)	1.801 (1.0252)	0.07890 017	NA	NA	2.418 (1.3212)	0.06728 618	NA	NA	1.623 (1.0399)	0.11855 64	1.849 (1.0312)	0.07288 724	NA	NA
Cigarette Pack Years	0.001 (0.002)	0.55620 48	0.001 (0.0021)	0.73209 49	0.025 (0.015)	0.09244 17	0.004 (0.0044)	0.35030 19	0.000 (0.0024)	0.95749 62	0.001 (0.0024)	0.52968 56	0.001 (0.0045)	0.89232 98
Tumor Grade (Ref= Localized)														
Regional by direct extension	0.627 (0.2756)	0.02290 213	0.564 (0.292)	0.05329 258	0.049 (1.1609)	0.96611 36	0.666 (0.4894)	0.17352 66	0.505 (0.3356)	0.13268 33	0.627 (0.3501)	0.07317 424	1.082 (0.5069)	0.03276 5
Regional to lymph nodes	0.981 (0.3044)	0.00126 432	0.971 (0.3209)	0.00249 059	-0.552 (1.7036)	0.74591 625	1.450 (0.5561)	0.00910 546	0.785 (0.3698)	0.03385 527	1.127 (0.3852)	0.00344 817	1.218 (0.5496)	0.02667 463
Regional by both direct extension and regional lymph nodes	0.813 (0.3033)	0.00732 828	0.778 (0.322)	0.01572 659	1.303 (1.2909)	0.31277 503	1.649 (0.5698)	0.00381 321	0.473 (0.3649)	0.19504 88	0.834 (0.3842)	0.02994 596	1.153 (0.5549)	0.03764 9
Unknown/ Unstageable	0.460 (0.2906)	0.11384 72	0.315 (0.313)	0.31468 58	0.345 (1.2546)	0.78335 678	1.132 (0.5254)	0.03114 618	0.176 (0.3602)	0.62567 46	0.341 (0.3736)	0.36088 48	1.067 (0.5217)	0.04086 074
Positive Nodes (Ref = All sentinel nodes examined are negative)														
Sentinel nodes are positive	0.280 (0.151)	0.06378 474	0.235 (0.1596)	0.14076 8	0.314 (0.6651)	0.63697 554	0.301 (0.2765)	0.27578 64	0.203 (0.1921)	0.28961 65	0.225 (0.1883)	0.23172 24	0.236 (0.2842)	0.40570 07
Other/ Unknown	1.241 (0.137)	1.2829E-19	1.275 (0.1433)	5.8584E-19	2.629 (0.8164)	0.00128 337	1.388 (0.2604)	9.8261E-08	1.213 (0.1694)	7.979E-13	1.133 (0.1638)	4.6122E-12	1.62 (0.2855)	1.4029E-08
Tumor size	0.003 (7e-04)	1.4003E-06	0.003 (7e-04)	6.3175E-06	0.001 (0.0092)	0.92388 607	0.003 (9e-04)	0.00159 761	0.005(0.0018)	0.00468 469	0.004 (0.002)	0.02653 452	0.003 (8e-04)	0.00056 374
Geographical Region (Ref = non-Appalachia)														
Appalachia	-0.044 (0.	0.67235 06	-0.043 (0.	0.68639 51	0.567 (0.70	0.42187 794	-0.169 (0.	0.39717 59	-0.089 (0.	0.47884 59	NA	NA	NA	NA

	1049)		1072)		59)		1992)		1263)					
Histology (Ref = Adenocarcinoma NOS)														
Papillary Adenocarcinoma	-0.096 (0.2048)	0.6376371	-0.03 (0.209)	0.8840294	-0.768 (1.288)	0.5512367	-0.598 (0.4155)	0.1500898	0.100 (0.2407)	0.676876	0.002 (0.2398)	0.9925568	0.028 (0.4319)	0.949838
Tubulovillous Adenocarcinoma	-0.651 (0.4259)	0.1261181	-0.818 (0.4583)	0.07412704	-1.651 (1.7298)	0.33993434	-1.692 (0.7761)	0.02922619	-0.584 (0.5183)	0.25972 73	-0.101 (0.5191)	0.8456566	-1.971 (0.8295)	0.01747958
Mucinous Adenocarcinoma	0.328 (0.1689)	0.05249475	0.347 (0.1721)	0.04370414	1.130 (1.2673)	0.37242751	-0.272 (0.365)	0.4554246	0.415 (0.1978)	0.0360406	0.453 (0.1842)	0.01398821	-0.277 (0.481)	0.5641568
Other/ Unspecified	-0.004 (0.1314)	0.9764196	0.047 (0.1359)	0.7315225	-1.625 (0.6508)	0.01254258	-0.450 (0.2844)	0.1134517	0.139 (0.1549)	0.37127 1	0.042 (0.1656)	0.8012593	0.040 (0.2313)	0.8631467

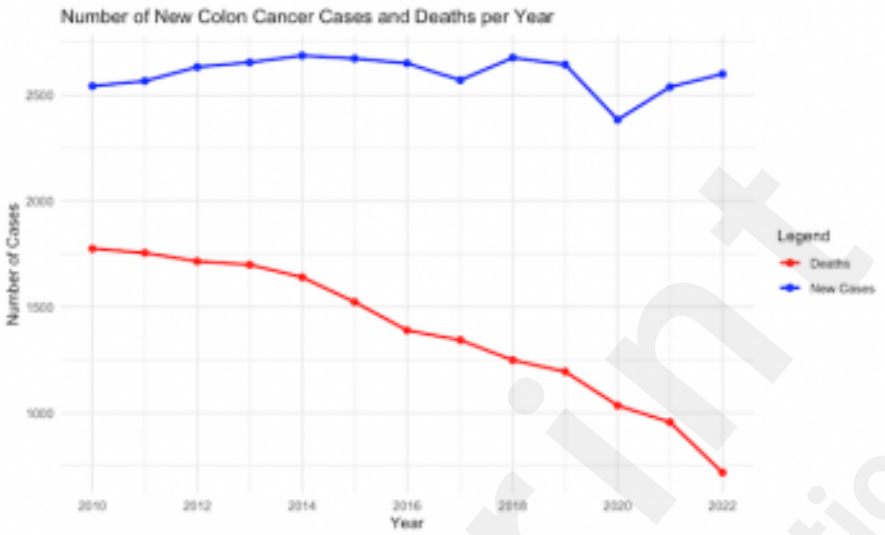
Footnote: Est (std): Estimated log hazard ratio from the Cox model (standard deviation), P-val: Adjusted p-value. Notice that the negative sign means lower hazard compared to the reference category as the estimated parameter of the covariate (risk factor) increases, the probability of experiencing death increases.

Supplementary Files

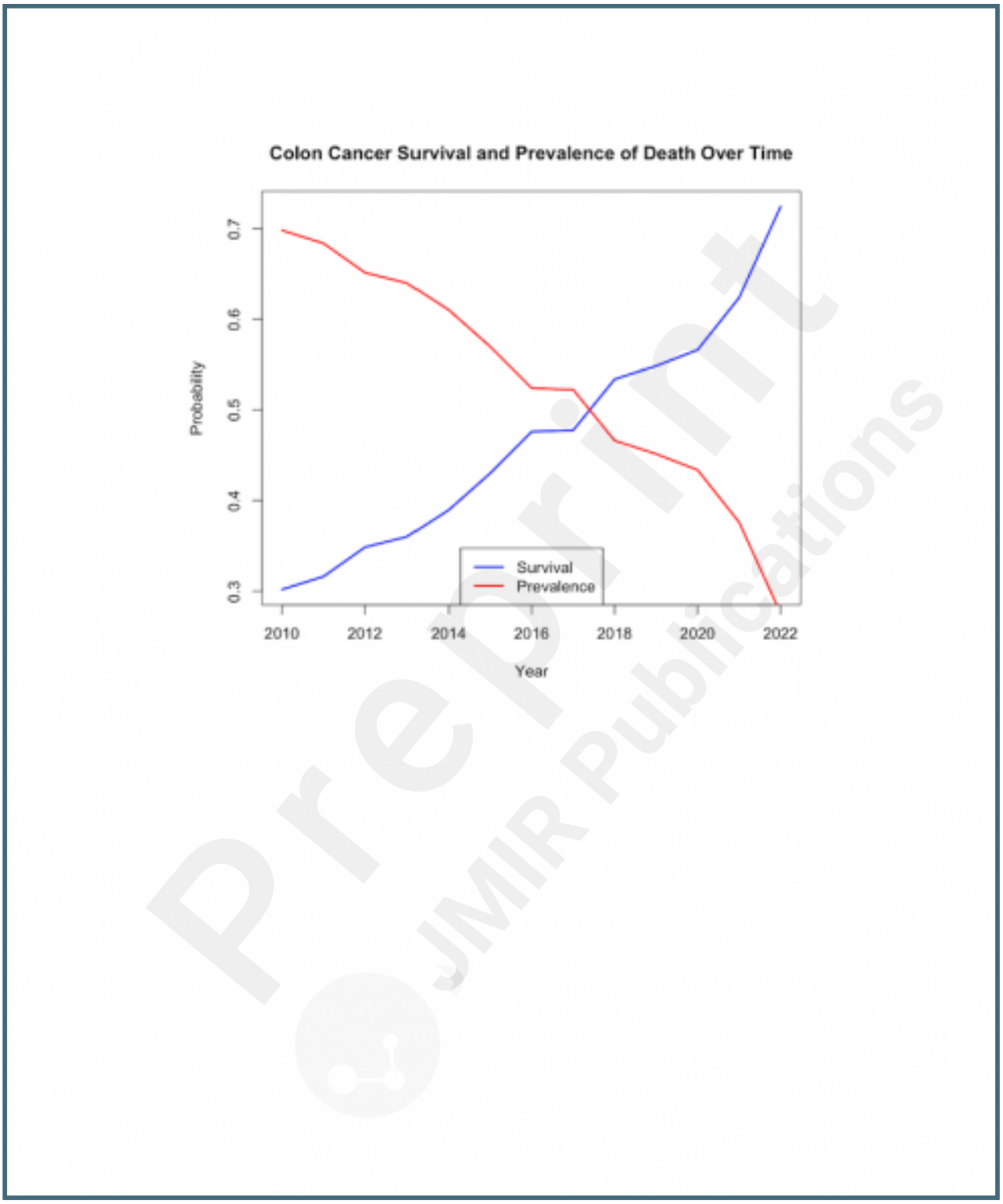
Untitled.
URL: http://asset.jmir.pub/assets/ef05da4ec59371aa89f088dcbf7f976f.docx

Figures

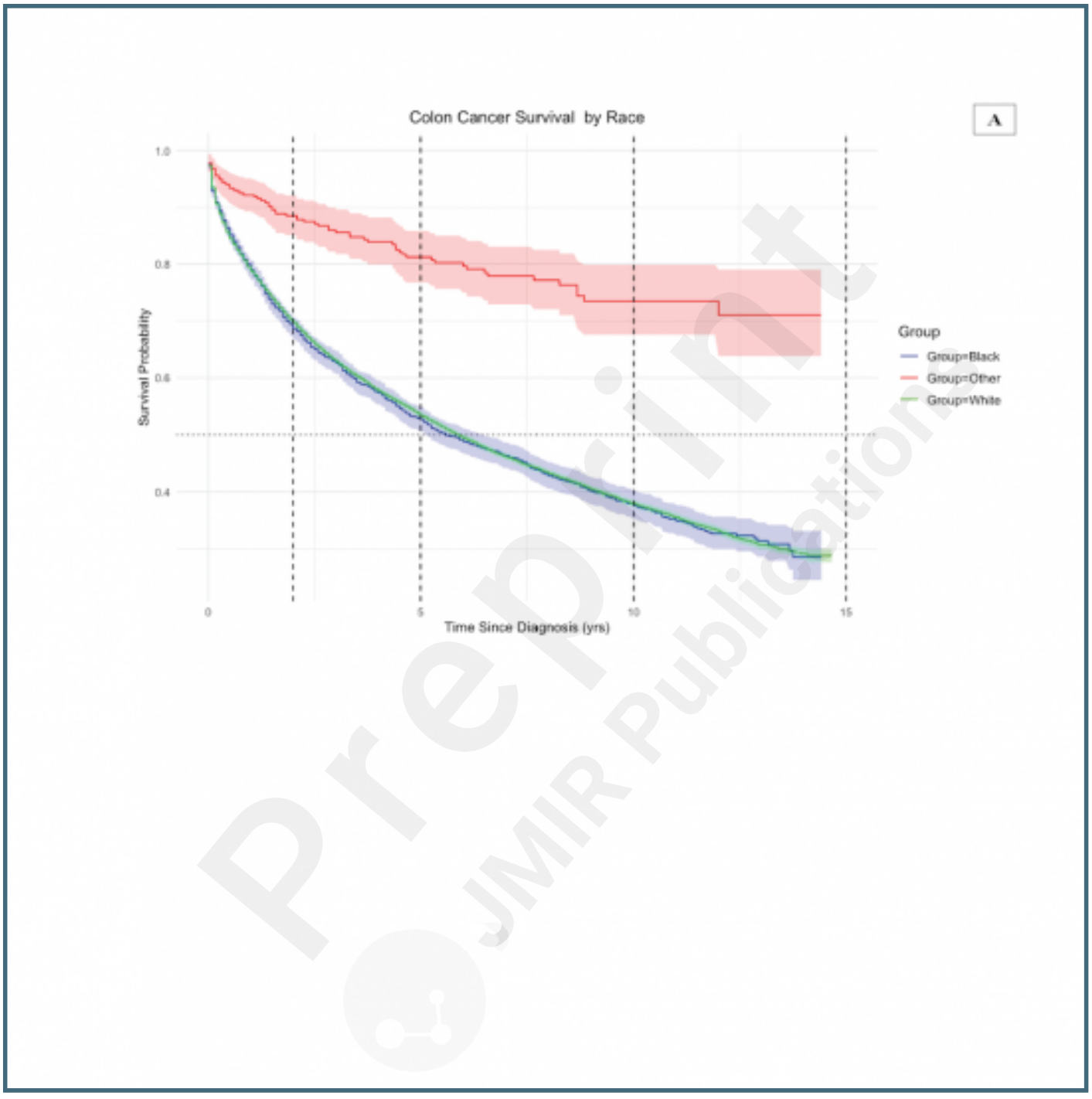
A comparison of new CRC cases and deaths across the study period.



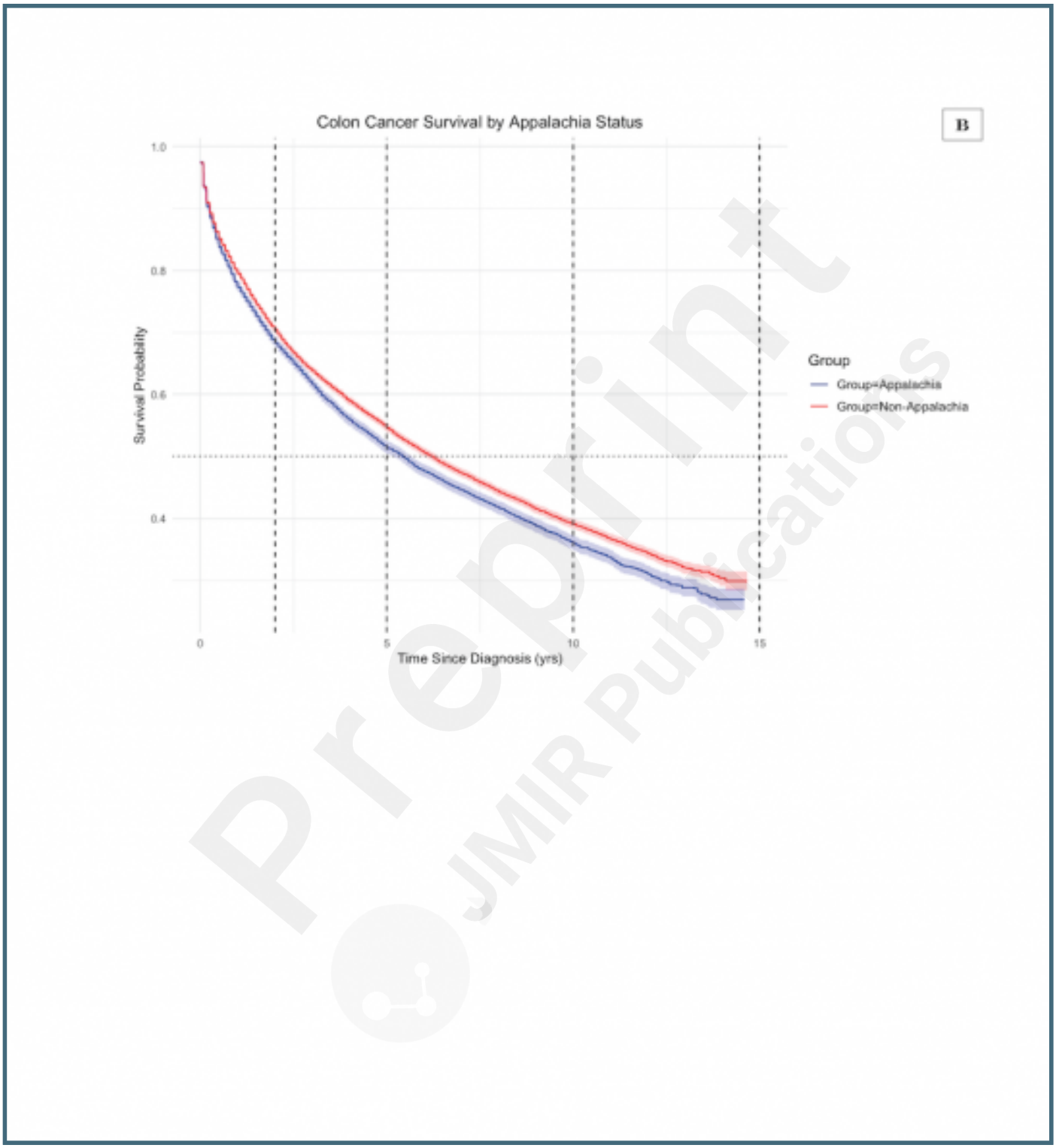
Understanding the CRC survival and prevalence across time.



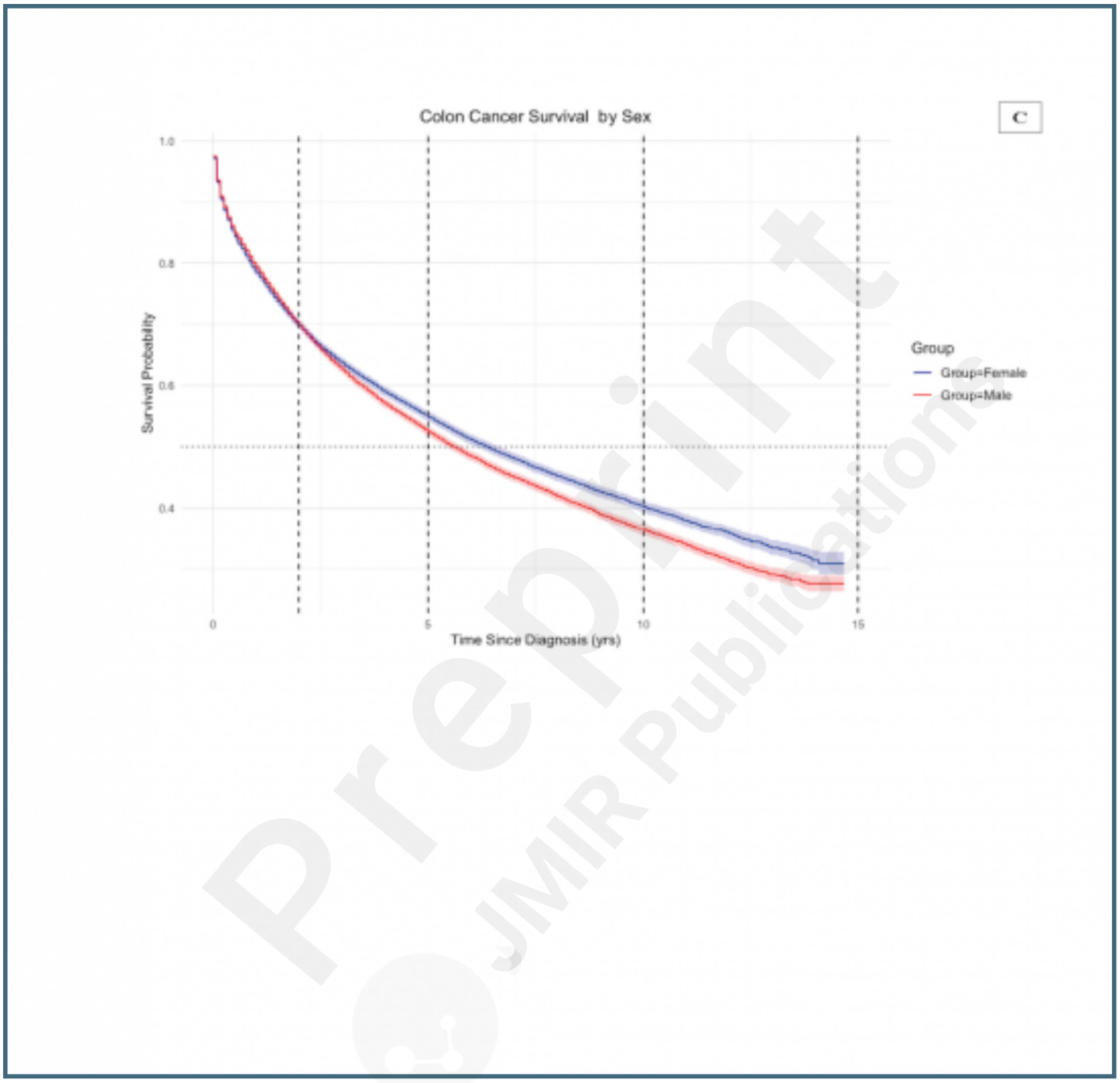
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



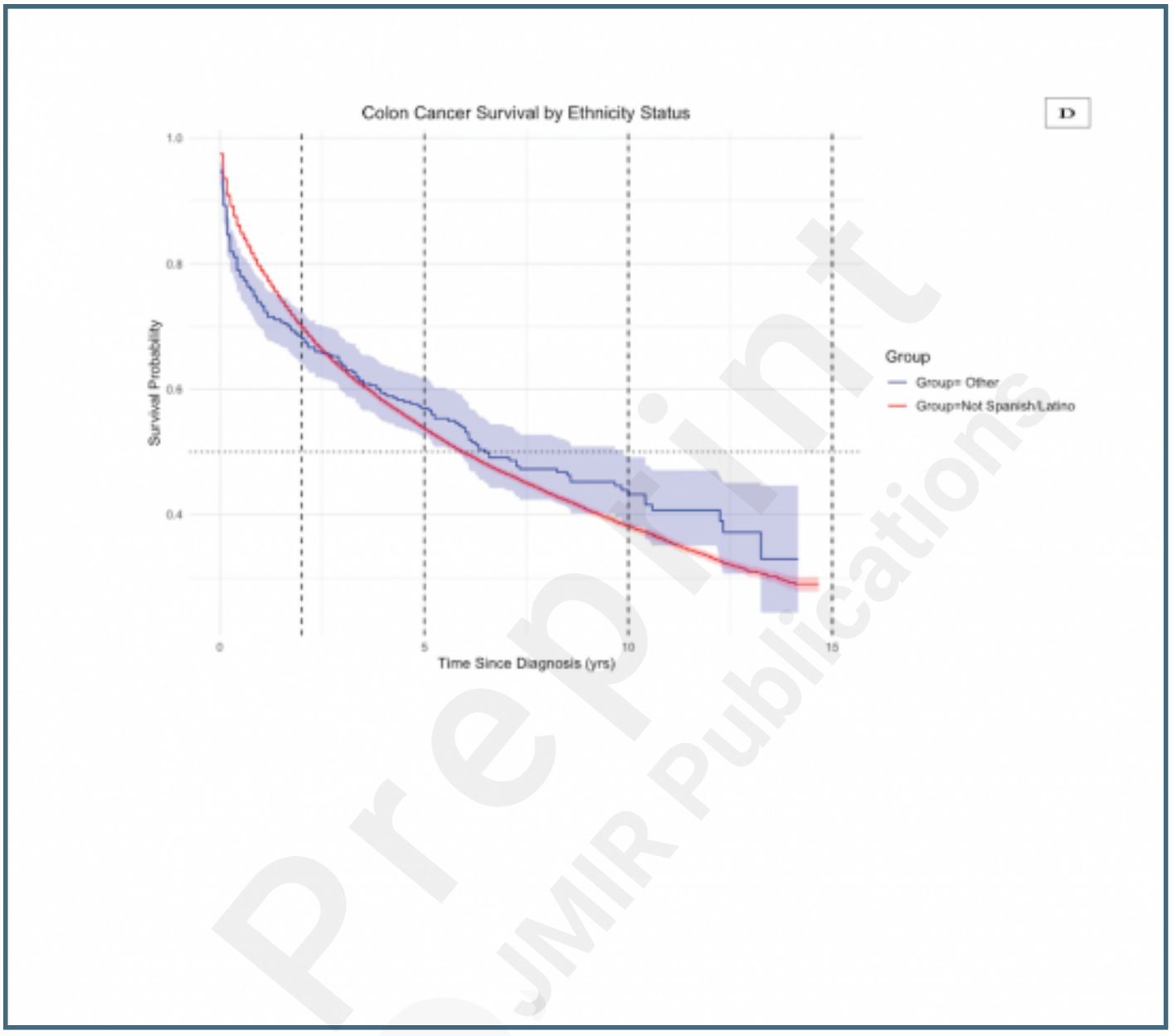
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



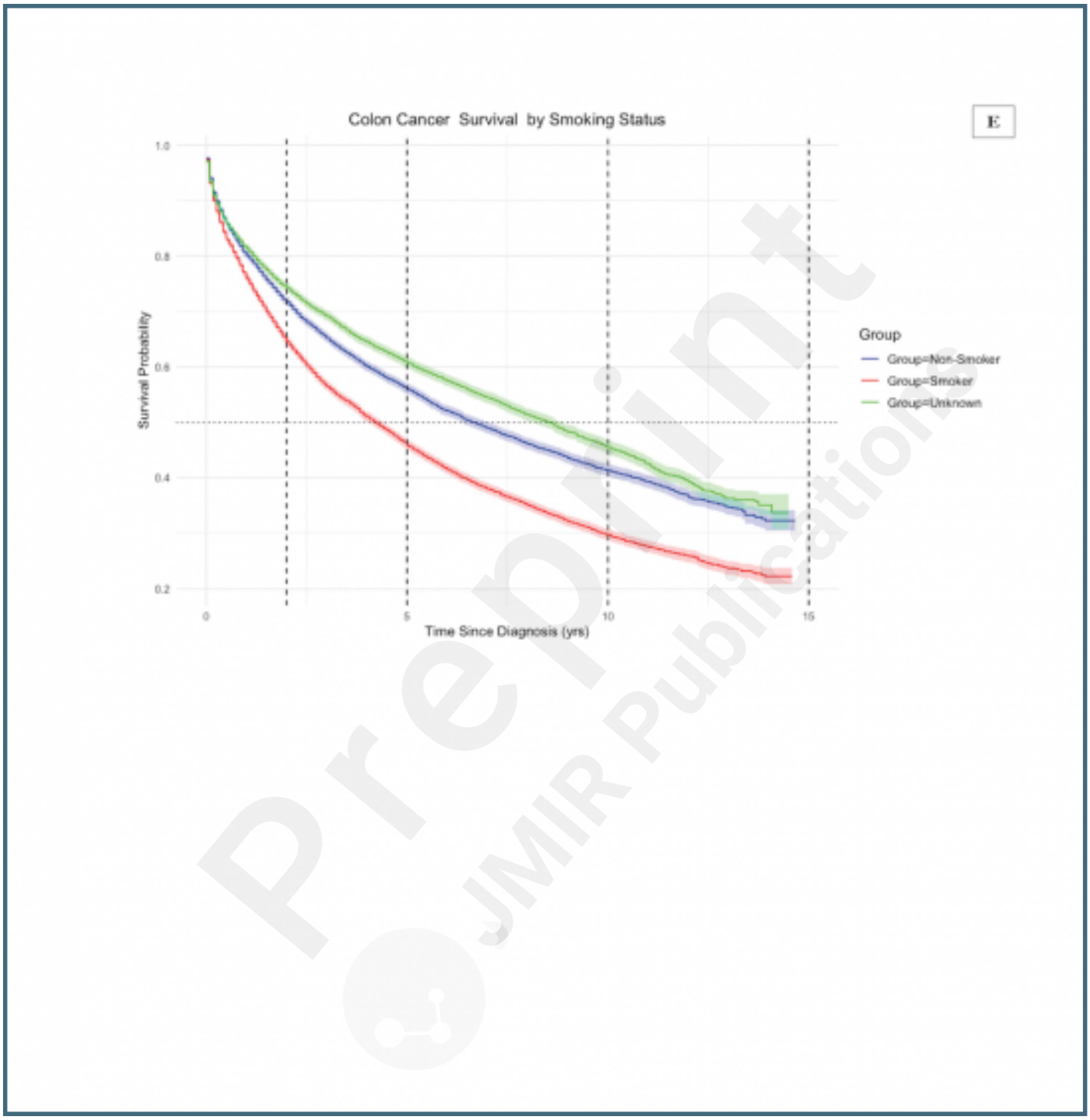
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



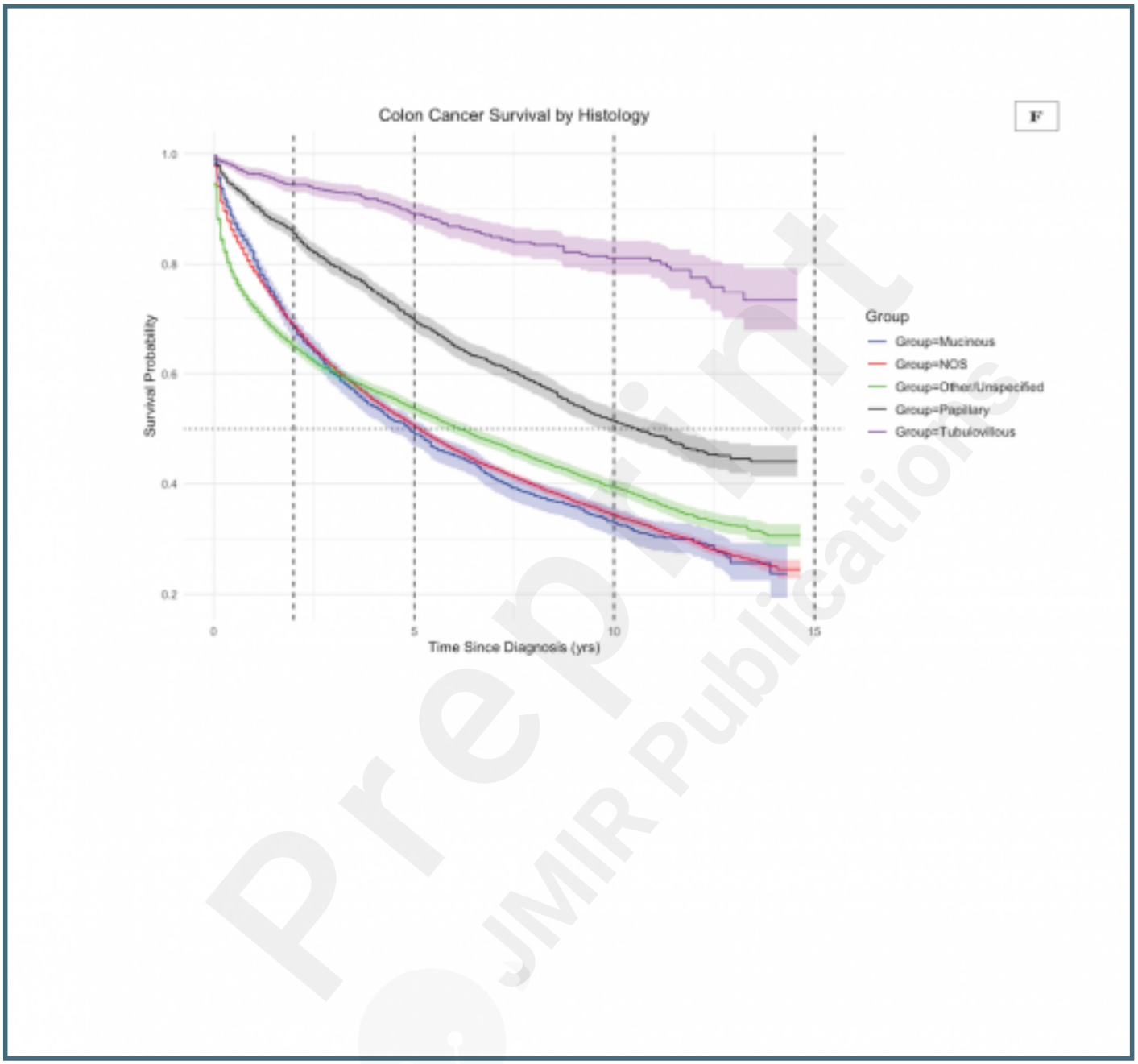
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



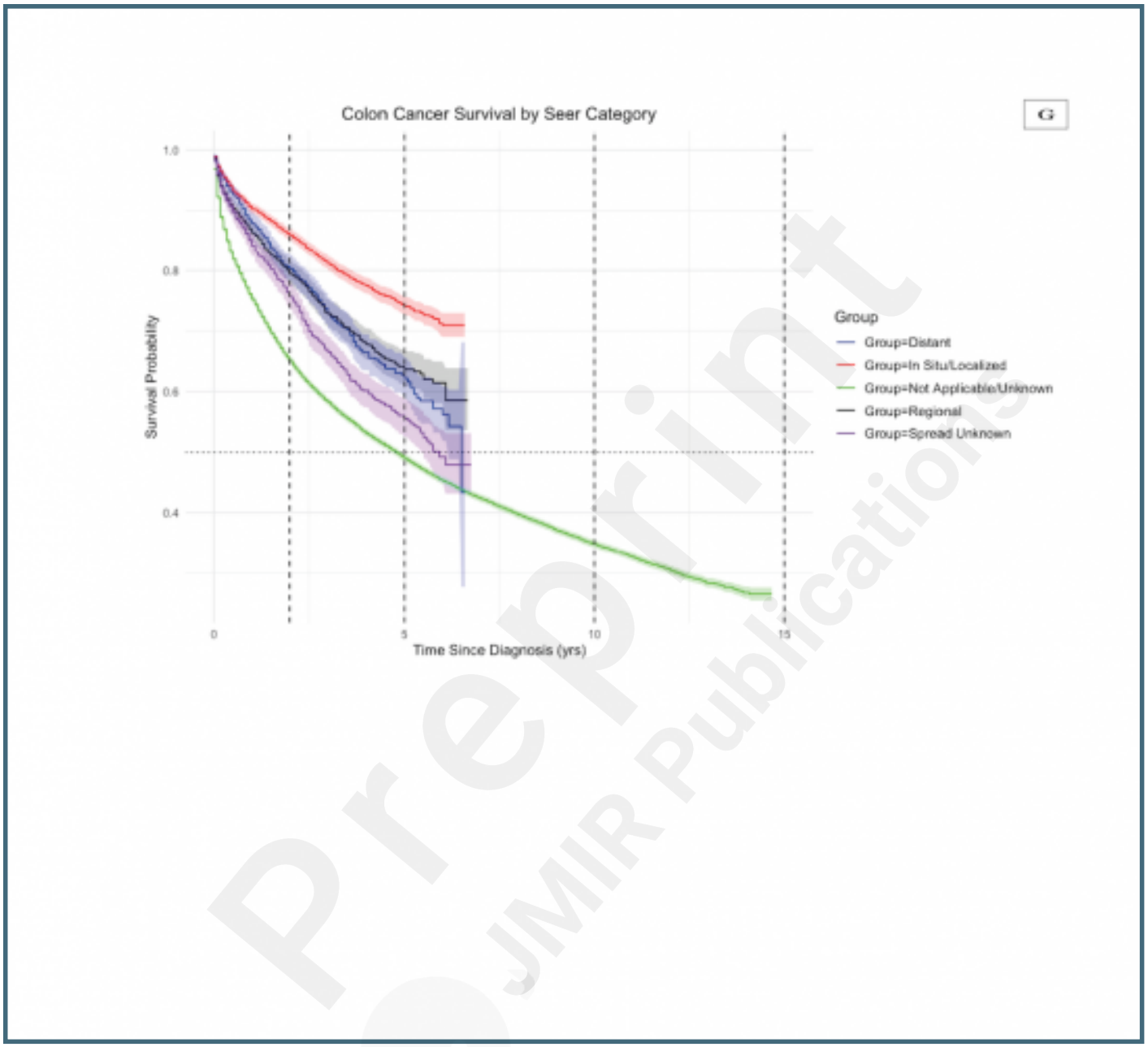
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



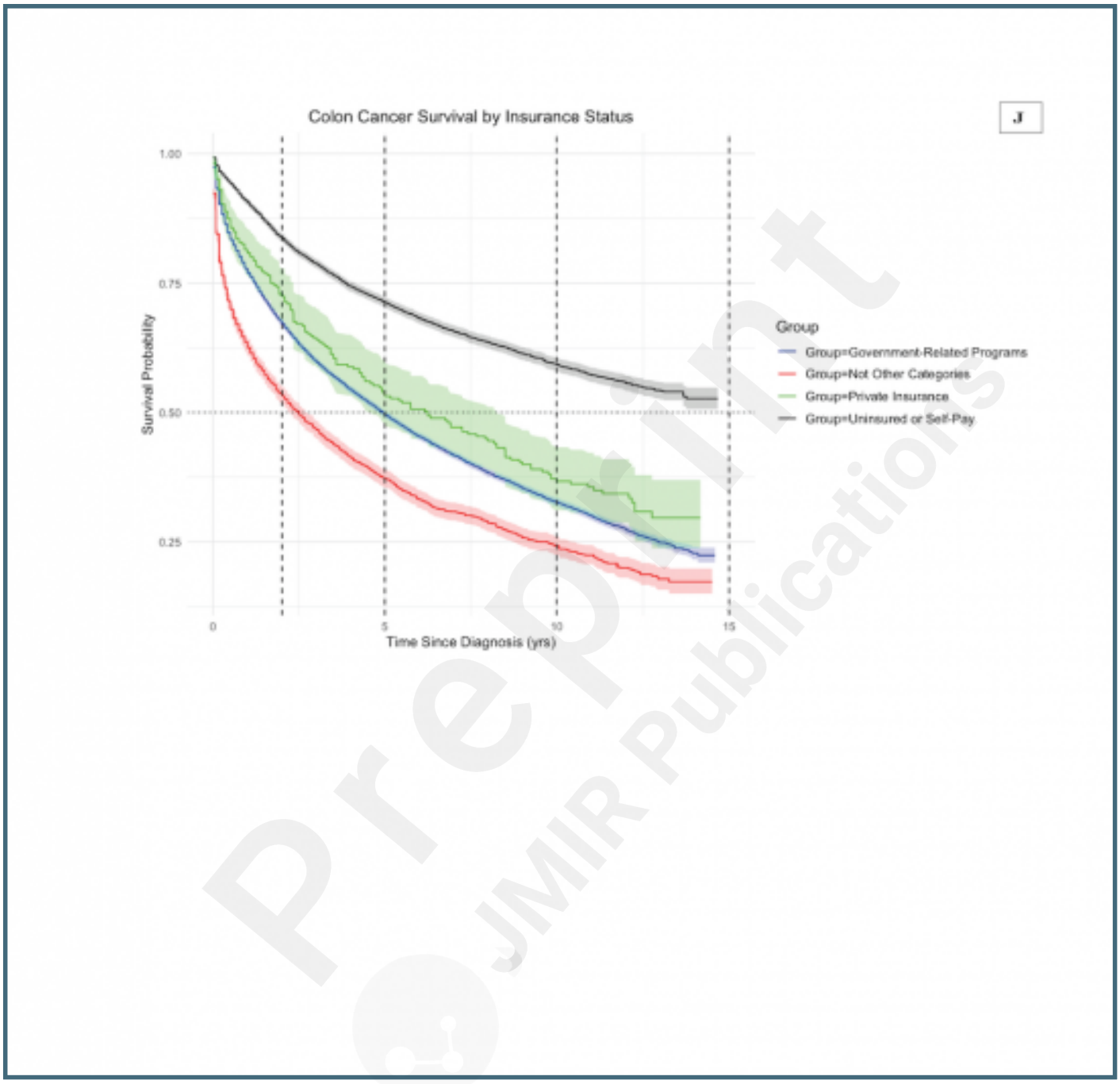
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



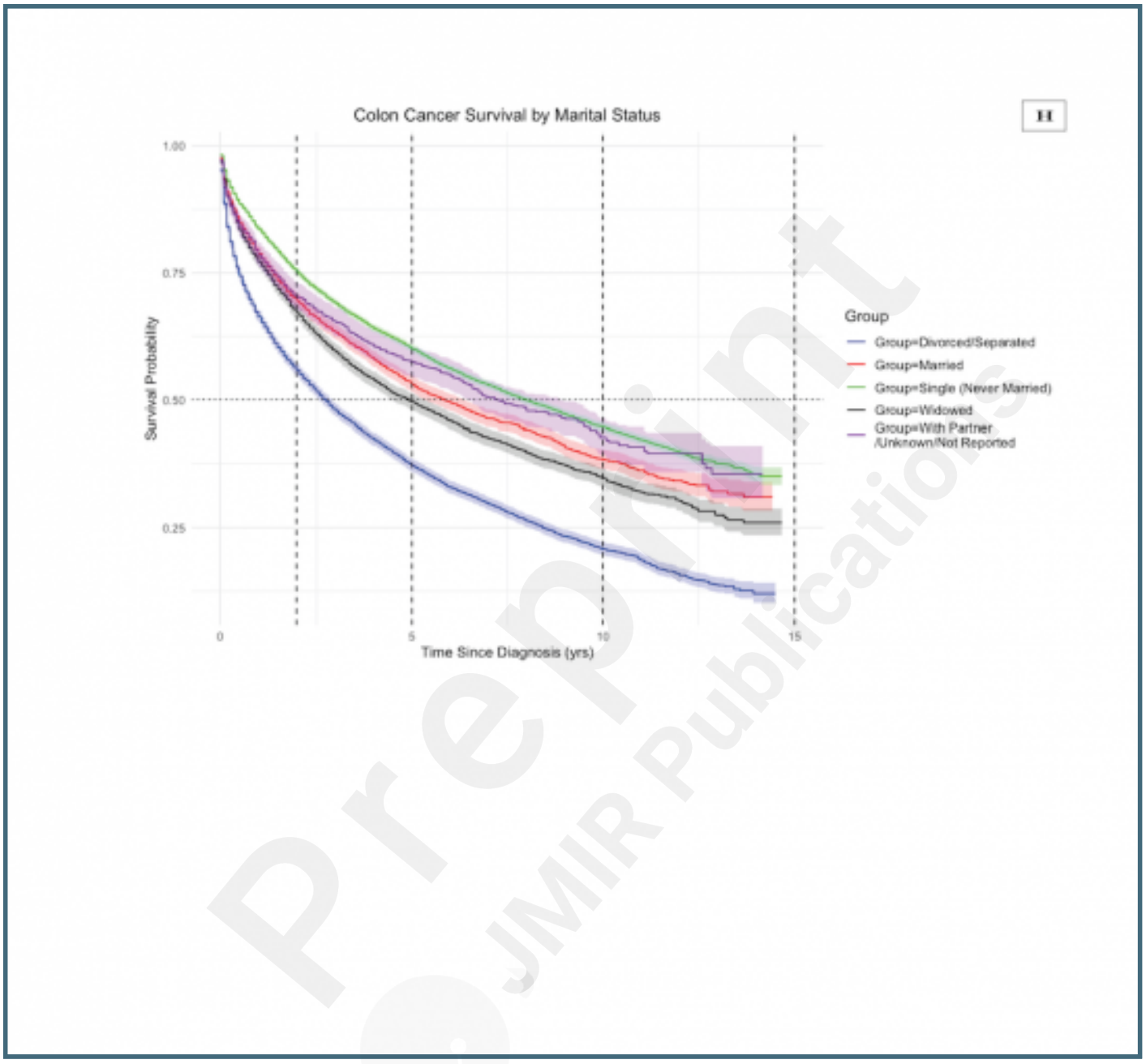
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



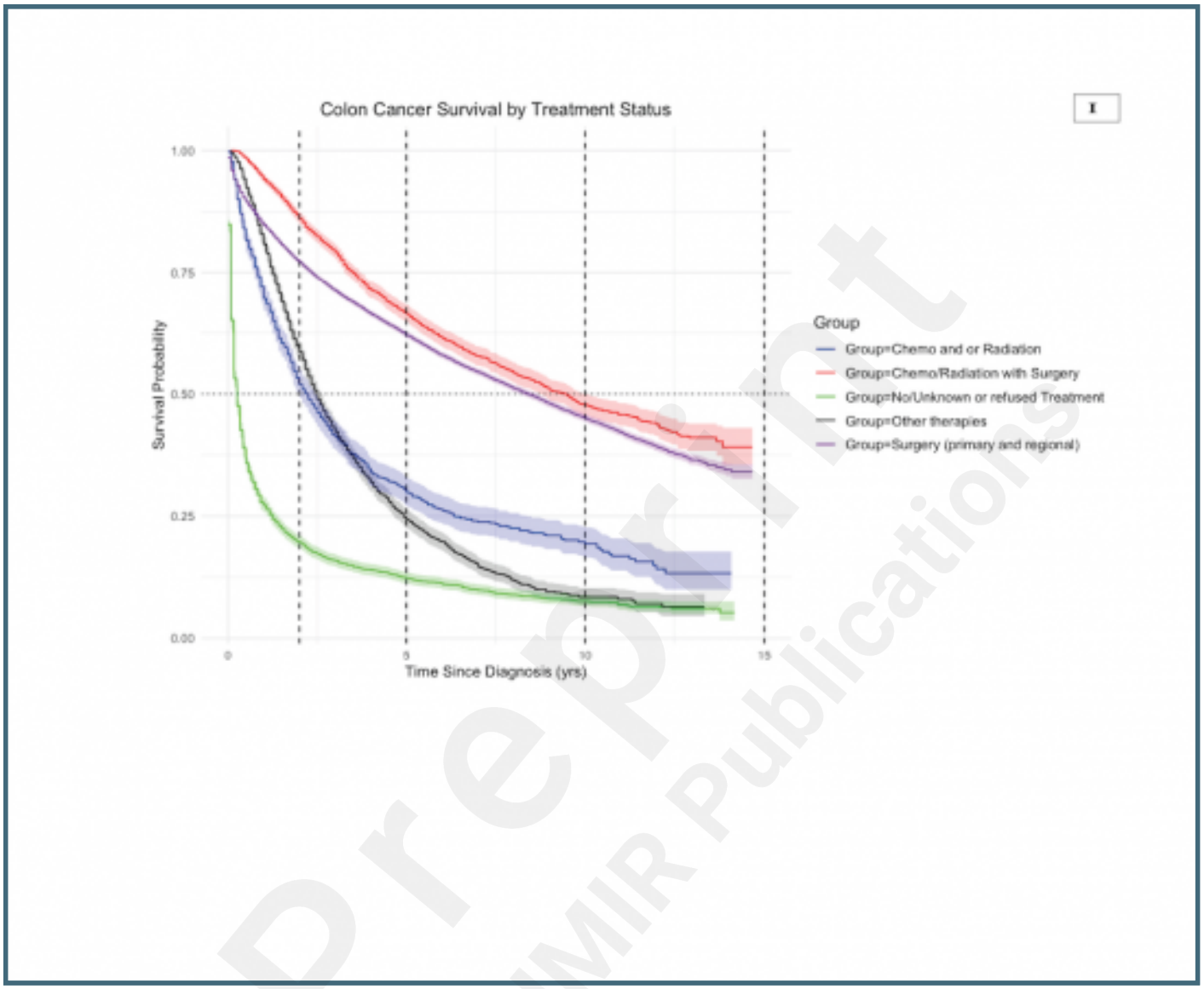
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



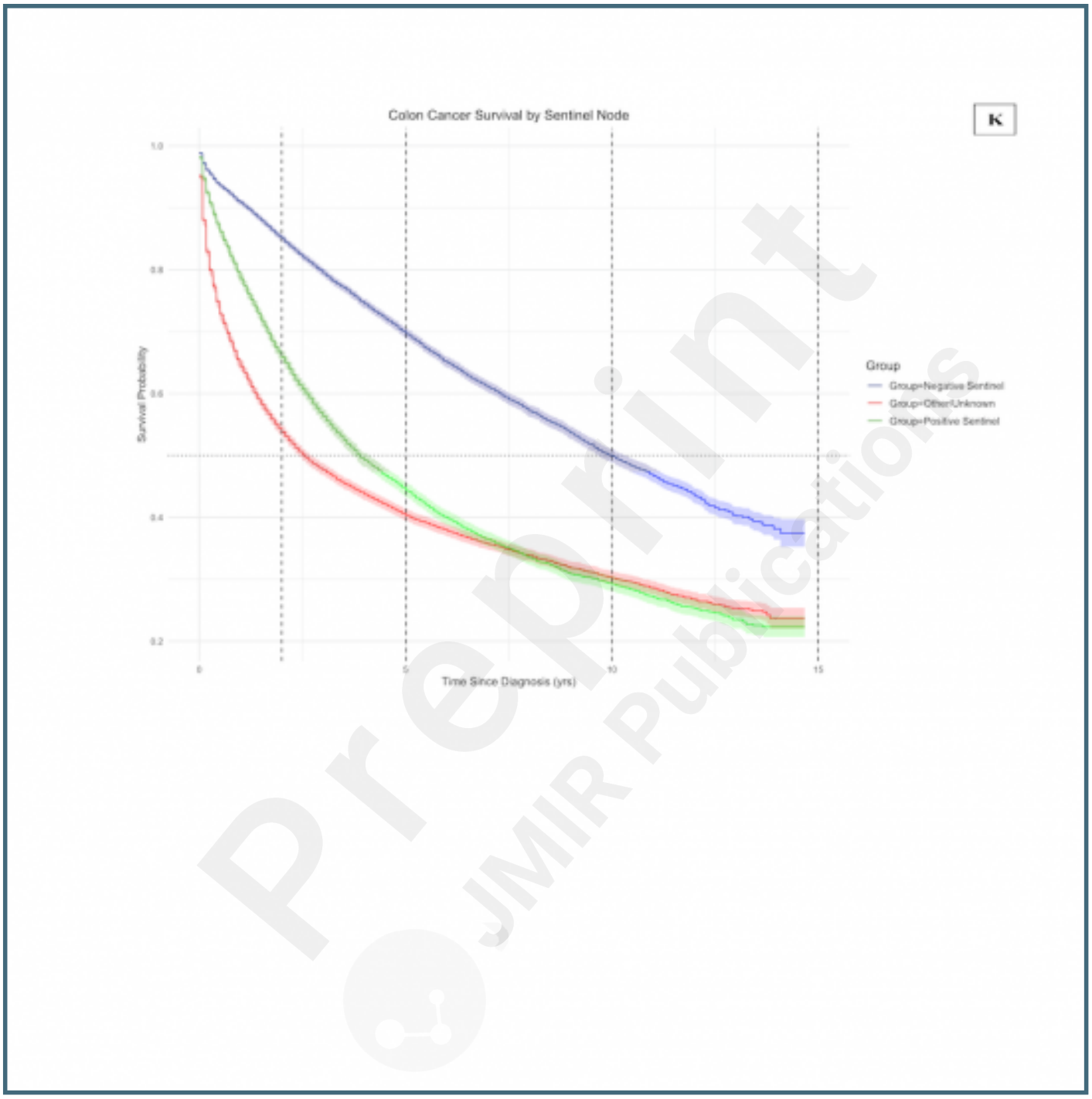
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



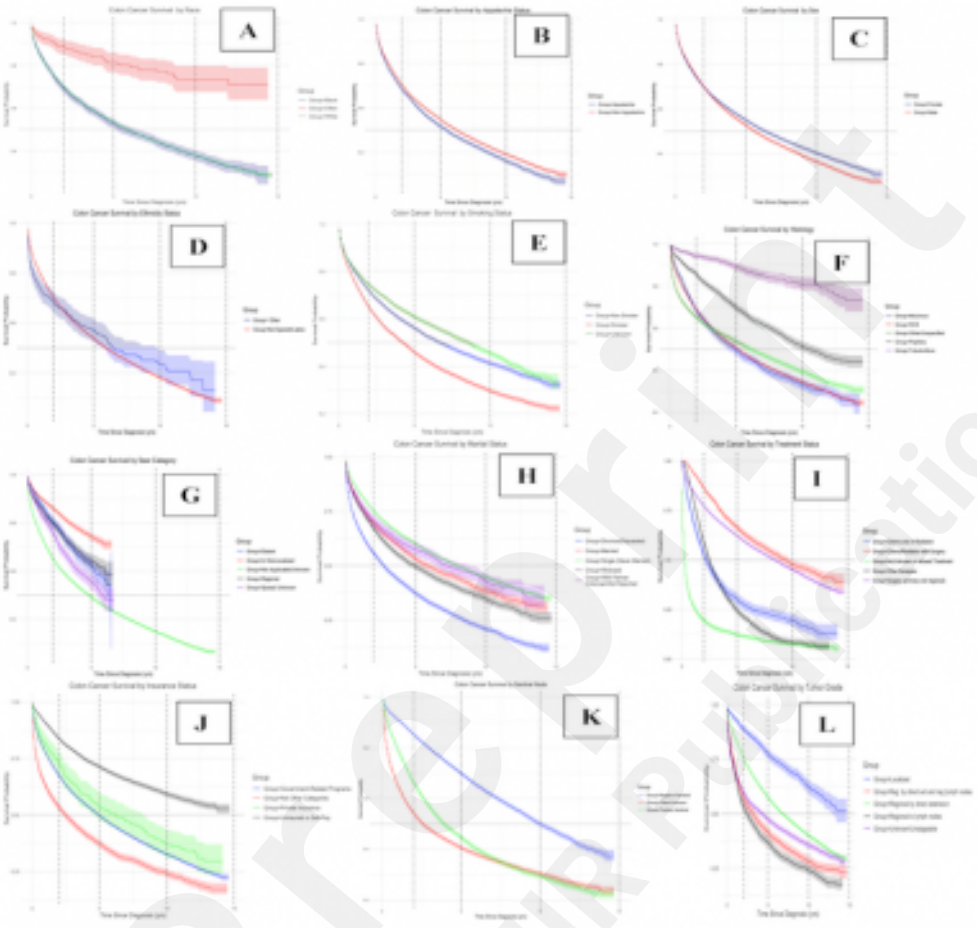
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



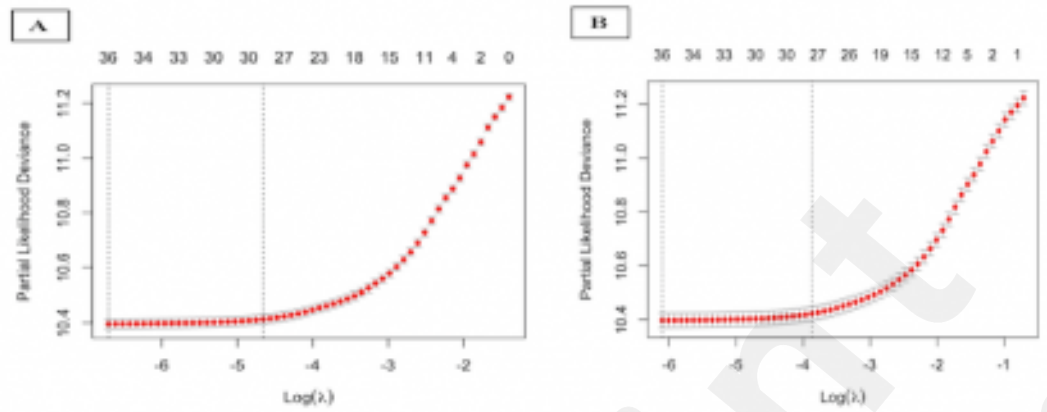
Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



Kaplan-Meier colon cancer survival curves comparison CRC cancer survival distribution by risk factors and time since diagnosis, with vertical dashes representing 2, 5, 10, and 15-year survival probabilities.



Cross validation for the Lasso (panel A) and the Elastic Net (panel B).



The overall ranking of risk factor importance and for each subgroup, as determined by the random survival forest model.