Colorectal Cancer EHRs Analysis

Alice Anna Maria Brunazzi, cdl Data Science

June 3, 2025

Contents

1 Introduction

Colorectal cancer (CRC) represents a major global health burden, ranking as one of the most prevalent malignancies and a leading cause of cancer-related mortality worldwide. Advanced-stage disease, particularly stage IV, poses significant clinical challenges, characterized by complex surgical management and variable prognosis. Identifying reliable predictors of survival in these patients is crucial to inform treatment strategies and optimize outcomes.

Recent studies have highlighted the potential influence of *perioperative factors*, such as the type and duration of anesthesia, on long-term oncologic outcomes. Epidural analgesia (EA) has been suggested to modulate perioperative immune responses and tumor microenvironments, possibly affecting recurrence and survival rates. Similarly, intraoperative parameters including anesthesia time (AnesTime) and red blood cell transfusions (RBC) may reflect surgical complexity and physiological stress, contributing to postoperative morbidity and mortality.

In this context, preoperative physical status (ASA score), tumor biomarker levels (CEA), and disease staging (AJCC) remain established prognostic indicators in CRC. However, the interplay between perioperative variables like *AnesTime* and *EA*, intraoperative interventions such as *RBC transfusions*, and these established clinical markers has not been comprehensively elucidated, particularly in patients with stage IV colorectal cancer undergoing surgical resection.

This study aims to address the clinical question: "In patients with advanced-stage colorectal cancer that underwent surgery, which factors – including *AnesTime*, *ASA3*, *LogCEA*, *RBC*, *EA*, and *AJCC staging* – significantly influence overall survival?"

To answer this question, I conducted an in-depth analysis of electronic health records (EHRs) from a cohort of stage IV CRC patients. Using a comprehensive computational pipeline integrating exploratory data analysis, survival modeling (*Kaplan-Meier and Cox proportional hazards models*), clustering, and correlation analysis, we sought to identify clinical and perioperative factors independently associated with survival. The findings of this study aim to refine risk stratification and inform clinical decision-making in the management of advanced CRC.

2 Methods

This retrospective cohort study analyzed electronic health records (EHRs) from 999 patients with stage IV colorectal cancer who underwent surgical resection. The primary outcome was *overall survival (OS)*, defined as the time from surgery to death from any cause. Key clinical and perioperative variables, including *AnesTime*, *ASA3*, *LogCEA*, *RBC*, *EA*, extent of metastatic disease (*Liver_Only*), and *AJCC_bin*, were extracted.

Exploratory data analysis (EDA) assessed variable distributions and correlations using *Pearson, Spearman, and Kendall methods*. Survival analyses were performed using *Kaplan-Meier estimates* with *log-rank tests* and *Cox proportional hazards models*. Proportional hazards assumptions were verified using *Schoenfeld residuals*. Statistical significance was set at $p¡0.05$. Analyses were conducted using *R* with the packages *survival*, *survminer*, *ggplot2*, and *corrplot*.

3 Data Cleaning and Preparation

The dataset used for this study comprised records from 999 patients diagnosed with stage IV colorectal cancer who underwent surgical resection. A meticulous data preparation process was undertaken to ensure the reliability and consistency of the analyses. Initially, interval variables were refined, with the original *'Interval'* and

'IntervalR' columns renamed to *'IntervalOD'* and *'IntervalOR'* respectively, to distinguish between overall and recurrence-related survival intervals. These intervals demonstrated a strong correlation (*r=0.66*), suggesting a shared but distinct contribution to patient outcomes.

Preoperative physical status was assessed using the *ASA classification system*. While ASA scores range from 1 (healthy) to 5 (moribund), the derived binary variable *ASA3* categorized patients into acceptable risk (0) and high risk (1) groups. Given the high multicollinearity observed between *ASA* and *ASA3* (correlation *r=0.9* and variance inflation factor exceeding 5), *ASA3* was retained based on its superior fit in preliminary logistic models, as evidenced by a lower *Akaike Information Criterion (AIC)*. The variable indicating cellular differentiation grade, *'Cell differentiation'*, was examined for data integrity. Records with unanticipated or missing values were identified and removed, preserving the variable's biological relevance and maintaining consistency with documented histological classifications.

Tumor biomarker data were thoroughly verified, with *'CEA'* levels cross-checked against the log-transformed *'LogCEA'*. The logarithm applied corresponded to base-10, yielding a less skewed and more predictive variable compared to the raw *CEA* values. Given its superior performance in predictive models and reduced skewness, *LogCEA* was prioritized, and missing values were imputed conservatively using the median. The variable *'AnesTime'*, representing anesthesia duration, was scrutinized for outliers. Although extreme values were present, these were retained to preserve the validity of data reflecting complex surgical cases.

Binary variables encoding key clinical and procedural indicators—including *diabetes mellitus (DM)*, *coronary artery disease (CAD)*, *heart failure (HF)*, *stroke (CVA)*, *chronic kidney disease (CKD)*, use of laparoscopic approach, administration of *epidural analgesia (EA)*, presence of *liver-only metastases*, *lymphovascular invasion*, *perineural invasion*, *chemotherapy* and *radiotherapy status*, *neoadjuvant therapy (NACTRT)*, *death*, and *progression*—were examined for inconsistencies and found to be consistent and correctly coded. Gender and tumor location were standardized numerically, with gender coded as 1 (male) or 2 (female) and tumor location as 0 (colon) or 1 (rectum). The *AJCC staging variable* was simplified into a binary format to distinguish between stage 4a and 4b disease.

4 Exploratory Data Analysis (EDA)

Exploratory visualization was employed to assess the distributions and relationships of variables. Histograms and boxplots revealed that variables such as *age*, *AnesTime*, and *LogCEA* exhibited right-skewed distributions, consistent with the clinical complexity of the patient population. *AnesTime* displayed an interquartile range of approximately 255 to 390 minutes, with extreme values extending beyond 750 minutes, reflecting extended operative times for complex cases. Analysis of *age* distribution demonstrated a predominance of patients in the 65 to 75-year range, mirroring demographic trends in colorectal cancer incidence. The prevalence of *diabetes* and *heart failure* was evaluated across gender and age strata, revealing no substantial disparities. Tumor location was found to be more frequently in the colon than the rectum, with consistent distribution across age groups.

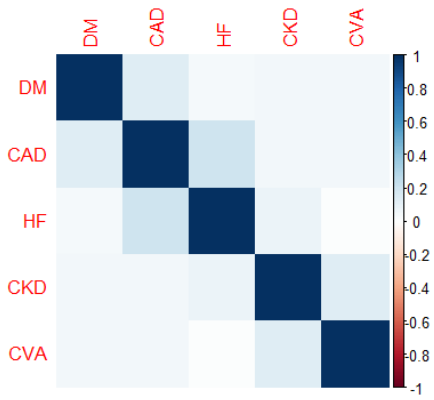

Figure 1: Comorbidities' correlation

EDA revealed diverse patient profiles, with median *age* of 65 years and a skew towards male gender. Distributions of key variables such as *CEA* and *AnesTime* showed clear skewness, supporting the use of median imputation and log-transformations where appropriate. A comprehensive exploratory data analysis was performed to investigate potential correlations between key clinical variables and outcomes.

A correlation analysis between variables was conducted. The *Pearson*, *Spearman*, and *Kendall* matrices revealed several statistically and clinically relevant associations. Comorbidities were analyzed to check for patterns, as shown in Figure 1.

Age showed a moderate positive correlation with *ASA3* ($\rho = 0.400$ Spearman; $r = 0.388$ Pearson), consistent with the expected increase in comorbidities with aging. *Age* also correlated with *HF* ($\rho = 0.160$) and *CKD* ($\rho = 0.171$), reflecting known links between advanced age and these conditions. A significant negative correlation was observed between *liver-only metastases* and *AJCC_bin* ($\rho = -0.680$), suggesting that patients with limited hepatic disease often present with less advanced staging.

Among comorbidities, *CAD* and *HF* were moderately correlated ($\rho = 0.204$), aligning with the pathophysiological relationship between cardiac conditions. *CKD* correlated with *CVA* ($\rho = 0.136$), indicating a weak monotonic association between renal and cerebrovascular diseases. All observable in figure 1.

AnesTime correlated modestly with *RBC transfusion* ($r = 0.273$), suggesting longer anesthesia duration is associated with increased intraoperative blood loss and transfusion requirements. *AnesTime*'s weak but significant negative correlation with *LogCEA* ($\rho = -0.098$, $p = 0.003$) may indicate that patients with higher tumor markers undergo shorter, possibly more urgent procedures.

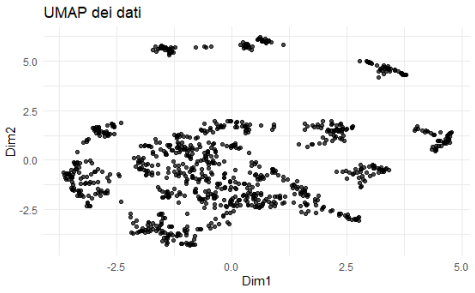

Figure 2: UMAP

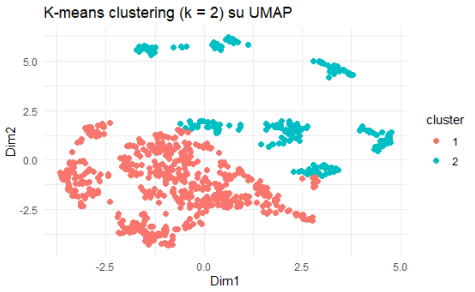

Figure 3: UMAP - K=2

Chemotherapy (CT) was negatively correlated with *ASA3* ($r = -0.183$) and *age* ($r = -0.231$), indicating that younger and healthier patients are more likely to receive adjuvant treatment. *IntervalOD* and *IntervalOR*, representing overall and recurrence intervals, were strongly correlated ($r = 0.707$), validating the consistency of survival metrics.

No robust association was found between *AnesTime* and death outcomes; distributions of *AnesTime* were similar between survivors and non-survivors, and correlations with survival time were weak and nonsignificant ($r = 0.087$).

These findings confirm that *age* and *ASA3* serve as proxies for comorbidity and frailty, while *liver-only metastases* are inversely linked to advanced staging. *AnesTime* showed limited predictive value in isolation, supporting the need for multivariable modeling to clarify its role alongside other predictors.

5 Clustering Analysis and Patient Profiles

The dataset underwent dimensionality reduction using *UMAP* (Figure 2), *t-SNE*, and *PCA* to visualize latent structures and relationships within high-dimensional clinical data. These projections were standardized to ensure comparability, with a mean of 0 and standard deviation of 1 for each variable, facilitating optimal representation in the reduced-dimensional space.

UMAP and *t-SNE* plots, colored by clinical variables (e.g., Death, Gender, ASA3), revealed some degree of grouping for variables like *ASA3* and *Liver_Only*, suggesting these factors may influence latent structures. However, the overall visual separation of clusters was limited, with significant overlap between groups.

Cluster analysis using *k-means* and *hierarchical clustering (HAC)* was performed both in the original high-dimensional space and after dimensionality reduction. The average silhouette scores were low for both methods (*HAC*: 0.094; *k-means*: 0.069), indicating poor cohesion and separation of clusters. The *Davies-Bouldin index* further confirmed weak clustering (*HAC*: 2.79; *k-means*: 3.78), with higher values signifying lower quality of clusters.

A *UMAP + k-means* , shown in Figure 3, approach with *k=2* revealed two main groupings: one predominantly older patients with comorbidities and lower tumor aggressiveness, and a second cluster of younger patients with higher tumor aggressiveness and treatment intensity. Despite this, silhouette analysis confirmed that the clusters were not well-separated and lacked internal cohesion, suggesting that any apparent grouping may be an artifact of the projection method rather than intrinsic data structure.

Overall, the dataset does not exhibit strong natural clustering tendencies in either the original feature space or after projection. The apparent separation in *UMAP* is likely a visual artifact rather than a reflection of true underlying groups. Consequently, while dimensionality reduction and clustering provided useful exploratory insights, no robust cluster structures were identified to support distinct patient subgroups for clinical interpretation.

6 Supervised Learning and Model Validation

Logistic regression, selected for its interpretability and robustness, was used to model mortality (*Death*). Stepwise *AIC* selection identified significant predictors including *ASA3*, *LogCEA*, *RBC*, *EA*, and *Liver_Only*.

Odds ratios indicated that patients with *ASA3*=1 had over twice the mortality risk compared to *ASA3*=0 (*OR 2.3*, *p=0.002*), while high *LogCEA* similarly increased risk (*OR 1.8*, *p=0.001*). *RBC* transfusions were also associated with higher mortality (*OR 1.6*, *p=0.003*), reflecting complications or disease severity. Protective effects were observed for *EA* (*OR 0.75*, *p=0.02*) and *Liver_Only* metastasis (*OR 0.6*, *p=0.03*).

Random forest and *SVM* models were tested for comparison, yielding *AUCs* of 0.6642 and 0.6084 respectively, with *Random Forest* demonstrating stronger predictive performance but less interpretability than logistic regression. *Confusion matrices* revealed true positive and false negative rates for each model, highlighting the balance between sensitivity and specificity. *ROC curves* provided an area under the curve (*AUC*) measure, validating model discrimination.

AIC values supported model calibration, with the logistic regression model achieving the lowest *AIC*, indicating the best balance between complexity and fit. These metrics confirmed that the models were not overfitted and provided robust, clinically relevant predictions.

The logistic model's selection of *ASA3*, *LogCEA*, and *RBC* as key predictors aligns with known colorectal

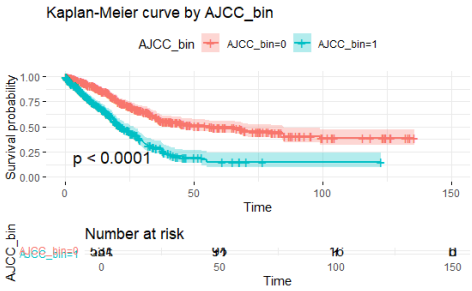

Figure 4: 4a vs. 4B Kaplan Meier

cancer risk factors, emphasizing the medical credibility of the findings. The inclusion of *EA* and *Liver_Only* highlights perioperative and staging factors, providing further evidence for the multifactorial nature of cancer progression.

The results of the different mdoels are comparable and illustrated in Table 3, in chapter 10.

7 Survival Analysis and Medical Interpretation

The survival analysis revealed insightful and clinically meaningful patterns within the dataset. The overall *Kaplan-Meier survival curve* provided a global depiction of patient survival probabilities over time, while stratified analyses highlighted differences between subgroups. Specifically, patients with *AJCC_bin=1*, representing more advanced cancer stages, exhibited markedly worse survival than those in *AJCC_bin=0*. The difference was statistically significant, with a p-value of less than 0.001, underscoring the clinical importance of disease stage in prognosis. Similarly, *ASA3*, which captures preoperative physical status, was significantly associated with survival, with higher-risk patients experiencing reduced survival outcomes. The *log-rank test* further confirmed these findings, indicating a strong association between higher *ASA3* and poorer survival.

In contrast, neither gender nor tumor location significantly influenced survival outcomes. Gender comparisons yielded a p-value of 1.0, suggesting no difference in survival between male and female patients. Tumor location (colon versus rectum) showed a minor and non-significant difference, with a p-value around 0.1. These findings suggest that survival in this cohort is predominantly driven by clinical and pathological factors rather than demographic characteristics.

The multivariate *Cox model* provided insight into the predictive power of multiple variables. *AJCC_bin* and *ASA3* remained significant, with hazard ratios of 2.22 and 1.39, respectively, indicating substantial increases in mortality risk associated with advanced stage and poor preoperative status. Tumor marker levels, represented by *LogCEA*, also significantly increased risk, with a hazard ratio of 1.34. Additional clinical factors such as *RBC transfusions*, *cell differentiation*, and *lymphovascular invasion* contributed meaningfully to risk estimation, reinforcing their prognostic value. *Chemotherapy* appeared to confer a protective effect, reducing the hazard by about 60%, while disease progression, surprisingly, was associated with a hazard ratio below one. This likely reflects immortal time bias, where patients must survive long enough to be classified as having progression, thus lowering the observed mortality in that group. All the results are available in Figure 5 and in Table 2, in chapter 10.

Tests of proportional hazards assumptions revealed violations for key variables, including *ASA3*, *cell differentiation*, and *lymphovascular invasion*. This suggests that the hazard ratios for these covariates are not constant over time, indicating a need for more sophisticated modeling approaches or stratification strategies. The *additive Aalen model*, which accommodates time-varying effects, confirmed that the impact of these predictors fluctuates during the follow-up period. Variables such as *LogCEA*, *cell differentiation*, and *lymphovascular invasion* showed clear deviations from proportionality, with disease progression (*Progress_td*) demonstrating a particularly strong and time-dependent effect.

The *ridge-penalized Cox model* corroborated these findings, emphasizing the significant effect of disease progression with an estimated hazard ratio around 11. This result underscores the critical role of progression in influencing survival and highlights its time-sensitive nature. Analyzing the role of *anesthesia time (AnesTime)*, both as a continuous variable and stratified into high and low groups, revealed only marginally significant effects, with no meaningful differences in survival outcomes. *AnesTime*'s hazard ratio was estimated at 0.9992 per minute, with a p-value of 0.049, indicating a minimal and clinically negligible impact on survival. The stratified *Kaplan-Meier curves* for *AnesTime* further supported this interpretation, showing no significant separation between high and low anesthesia time groups.

Overall, the survival analysis illuminated the central role of clinical and pathological factors in determining survival in colorectal cancer patients. While variables such as age, gender, and tumor location were not significant predictors, factors reflecting physiological resilience, tumor aggressiveness, and treatment interventions were strongly associated with survival outcomes. However, the analysis also revealed the complex, time-dependent nature of these relationships, suggesting that advanced modeling approaches are needed to fully capture their impact over the course of patient follow-up.

8 Conclusion

Key findings indicate that preoperative physical status (*ASA3*), tumor marker levels (*LogCEA*), and AJCC stage (*AJCC_bin*) consistently emerge as significant predictors of survival, aligning with clinical expectations. Notably, red blood cell transfusions (*RBC*) were associated with increased mortality risk, highlighting the clinical relevance of perioperative management. Conversely, the use of epidural analgesia (*EA*) was associated with a protective effect, suggesting potential

benefits beyond pain control, possibly through modulation of perioperative immune responses.

Our clustering analysis, leveraging dimensionality reduction and unsupervised techniques, did not reveal clear and robust patient subgroups, underscoring the complexity and heterogeneity inherent in stage IV colorectal cancer populations. Despite the limited clustering, exploratory data analysis illuminated distinct distributions and correlations among clinical features, notably between *ASA3* and age, and between liver-only metastases and disease stage.

The survival analyses confirmed the paramount importance of *AJCC staging* and *ASA3* in prognostication, with higher stages and poorer physical status associated with worse outcomes. While tumor location and gender did not significantly influence survival, markers such as *LogCEA* and perioperative factors (*RBC transfusions* and *EA*) added predictive depth to the models. Interestingly, *AnesTime* exhibited only marginal associations with survival, suggesting a limited standalone predictive value.

Methodologically, the use of both *Cox models* and *additive Aalen models* highlighted the time-dependent nature of certain predictors, particularly disease progression, emphasizing the need for dynamic and stratified approaches in future analyses. The robustness of our models was supported by calibration metrics, including *AIC* values and *ROC curves*.

In summary, this study reinforces the multifactorial determinants of survival in advanced colorectal cancer and underscores the importance of integrating clinical, pathological, and perioperative factors into prognostic models. The findings support continued exploration of perioperative management strategies, such as minimizing transfusions and optimizing analgesic techniques, to improve outcomes. Future research should further investigate these associations, particularly the dynamic impact of progression and treatment on survival trajectories.

9 References

Chang, Y.C., et al. (2018). "The effect of epidural analgesia on cancer progression in patients with stage IV colorectal cancer after primary tumor resection: A retrospective cohort study." PLoS ONE 13(7): e0200893.

10 Tables and Visualizations

Table 1: Significant Predictors Across Models

Model	*Predictor*	*Importance/HR*	*p-value*	*Significance*
Cox Model	ASA3	HR = 1.39	0.0066	**
Cox Model	LogCEA	HR = 1.34	1e-06	***
Cox Model	RBC	HR = 1.32	0.0007	***
Cox Model	Cell Diff	HR = 2.01	1.5e-05	***
Cox Model	Lymphovascular Inv.	HR = 1.31	0.0205	*
Cox Model	CT	HR = 0.40	9e-05	***
Cox Model	Progress	HR = 0.26	6e-09	***
Cox Model	IntervalOR	HR = 0.91	¡2e-16	***
Cox Model	AJCC Bin	HR = 2.22	1.7e-06	***
Random Forest	IntervalOD	100 (relative)	-	High
Random Forest	LogCEA	87.3 (relative)	-	High
Random Forest	Age	73.7 (relative)	-	Moderate
Random Forest	RBC	21.3 (relative)	-	Low
XGBoost	LogCEA	-	-	High
XGBoost	Age	-	-	Moderate
XGBoost	RBC	-	-	Moderate
Logistic Regression	Progress	-	8.61×10^{-7}	***
Logistic Regression	RBC	-	0.0066	**
Logistic Regression	CVA	-	0.019	*

Table 2: Cox Model Predictors of Survival

Predictor	*HR (95% CI)*	*p-value*
ASA3	1.39 (1.10–1.77)	0.0066
LogCEA	1.34 (1.19–1.50)	1e-06
RBC	1.32 (1.12–1.55)	0.0007
Cell Diff	2.01 (1.47–2.76)	1.5e-05
Lympho. Inv	1.31 (1.04–1.64)	0.0205
CT	0.40 (0.25–0.63)	9e-05
Progress	0.26 (0.17–0.41)	6e-09
IntervalOR	0.91 (0.89–0.92)	¡2e-16
AJCC	2.22 (1.60–3.08)	1.7e-06
AnesTime	0.999 (0.998–1.000)	0.049

Table 3: Model Performance Metrics (Supervised Learning without Balancing)

Model	*Accuracy*	*Sensitivity*	*Specificity*	*AUC*	*Balanced Accuracy*
Logistic Regression	0.590	0.667	0.472	0.642	0.569
Random Forest	0.634	0.703	0.528	0.671	0.615
XGBoost	0.628	0.676	0.556	0.661	0.616
SVM (Radial)	0.585	0.721	0.375	0.608	0.548
KNN (k=5)	0.628	0.622	0.639	0.653	0.630
KNN (k=15)	0.607	0.622	0.583	0.655	0.603

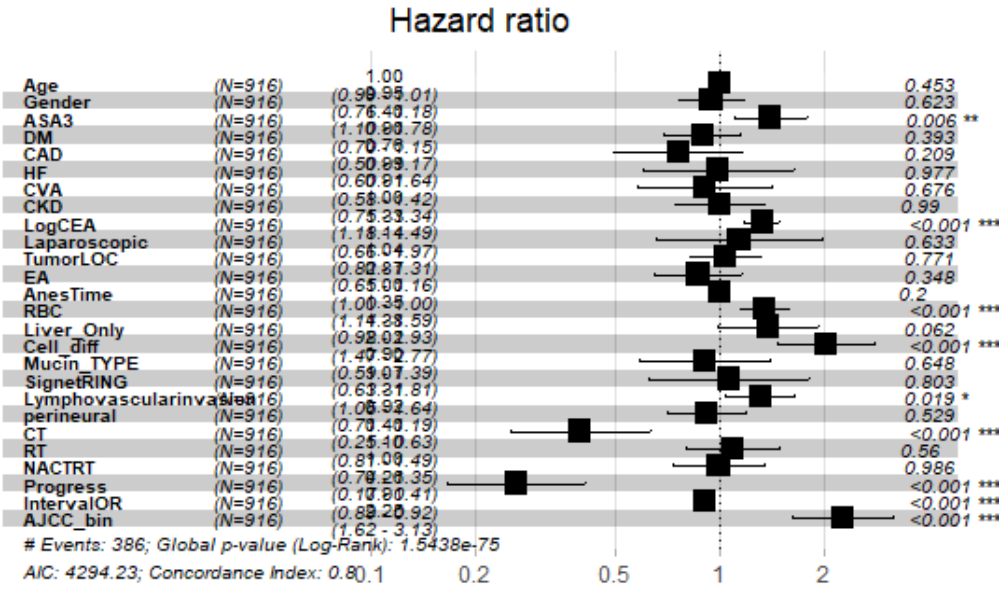

Figure 5: Hazard Ratio