Associations of Healthcare Costs and Early Detection

with Colorectal Cancer Mortality in Canada

Yulin Yuan, Ruohan Sun, Tony Lee

April 9, 2025

Introduction

Study Objective

The objective of this study is to evaluate how healthcare expenditures and early detection of colorectal cancer are associated with patient mortality outcomes in Canada. Our goal is to explore whether patients whose cancer is detected at an early stage, and those who bear different levels of healthcare costs, exhibit significantly different mortality rates. Through statistical analysis of available data, we will quantify the strength of these associations and assess their significance. We can build on the findings of this study to inform healthcare decisions, such as emphasizing effective screening programs and optimizing the use of healthcare resources to improve patient outcomes.

Dataset Overview

The dataset we decide to conduct our analysis of Mortality status (Alive/Dead) was sourced from Kaggle, which the dataset name is "Colorectal Cancer Global Dataset & Predictions". For the objective of this project, we only considered the data from Canada, which includes patient demographics, lifestyle risks, medical history, cancer stage, treatment types, survival chances, and healthcare costs. Key variables of we are going to investigate as follows:

Response Variable

- **Mortality Status**
  Status of patients' mortality, coded as `Yes` or `No`.

Explanatory Variables

- **Alcohol Consumption**
  Status of patient's alcohol consumption, coded as `Yes` or `No`.
- **Age**
  Patient's age in years.
- **Cancer Stage**
  Diagnosis stage of cancer, coded as `Localized`, `Regional`, or `Metastatic`.
- **Diabetes**
  Status of patient's diabetes, coded as `Yes` or `No`.
- **Diet Risk**
  Level of dietary risk, coded as `Low`, `Moderate`, or `High`.
- **Early Detection**
  Detection of colorectal cancer at an early stage, coded as `Yes` or `No`.

- **Family History**
  Family history of colorectal cancer, coded as `Yes` or `No`.

- **Genetic Mutation**
  Presence of genetic mutations for colorectal cancer, coded as `Yes` or `No`.

- **Gender**
  Gender of the patient, coded as `Male` or `Female`.

- **Healthcare Cost**
  Estimated healthcare expenditure per patient (in 1,000s of $).

- **Inflammatory Bowel Disease**
  Presence of inflammatory bowel disease, coded as `Yes` or `No`.

- **Insurance Status**
  Health insurance coverage, coded as `Insured` or `Uninsured`.

- **Obesity BMI**
  BMI classification, coded as `Normal`, `Overweight`, or `Obese`.

- **Physical Activity**
  Level of physical activity, coded as `Low`, `Moderate`, or `High`.

- **Screening History**
  Cancer screening history, coded as `Regular`, `Irregular`, or `Never`.

- **Smoking History**
  Patient's smoking history, coded as `Yes` or `No`.

- **Treatment Type**
  Type of treatment received, coded as `Surgery`, `Chemotherapy`, `Radiotherapy`, or `Combination`.

- **Tumor Size**
  Colorectal tumor size in millimeters (mm).

- **Urban or Rural**
  Patient's area of residence, coded as `Urban` or `Rural`.

Motivation

Colorectal cancer is one of the most common cancers in Canada, and it is predicted that rectal cancer will be the fourth most common cancer in Canada by 2024, making it critical to understand the factors that influence mortality (Du Cancer, 2024). Analyzing mortality in relation to healthcare costs and early detection is crucial because it addresses issues at the intersection of patient care and healthcare policy. It is widely recognized that the earlier a cancer is detected, the higher the chance of survival, so examining the impact of early detection of rectal cancer on mortality with actual patient data can help quantify the benefits of screening and early diagnosis efforts in Canada. Similarly, health care expenditures per patient can broadly reflect the intensity or quality of treatment received. Examining the relationship between healthcare expenditure and survival can help determine whether resources could be used more efficiently. By focusing on Canadian patients, we eliminate cross-national differences in healthcare systems, thereby providing a clearer understanding of the role of these factors within the context of a single-country healthcare system.

Analysis

Exploratory Data Analysis (EDA)

Balance of Response Variable

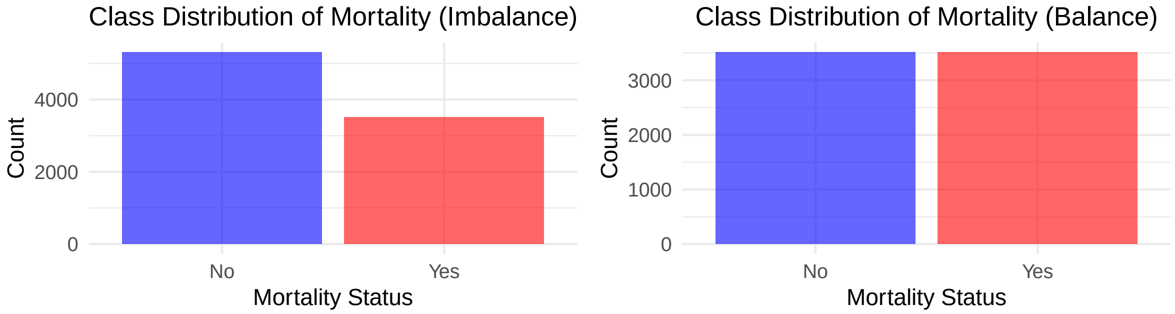

Figure 1: Class Distribution of the Mortality Variable Before and After Balancing

Figure 1 illustrates the distribution before and after balancing the mortality status. Many more patients were alive than dead before balancing the data, which could bias any predictive model. To address this issue, we created a balanced subset by randomly under-sampling Canadian patients with different mortality statuses, producing the same number of "Yes" and "No" patients. This step is critical to ensure that unbalanced mortality outcomes do not compromise subsequent modeling.

Continuous Variables

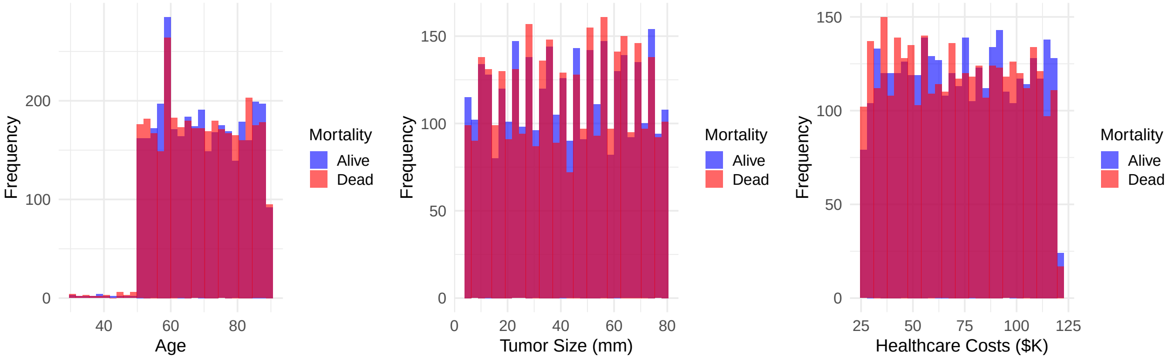

Figure 2: Distribution of continuous variables by mortality status

Next, we examined continuous variables (age, tumor size, and healthcare costs) that may be associated with mortality status. Figure 2 shows the distribution of these variables between living and dead patients. We found that patients who died tended to be older, had larger tumor sizes, and that medical costs were typically higher for deceased patients compared to the living. The distribution of cost variables (in thousands of dollars) shows a heavier upper tail for the deceased group, suggesting that higher medical expenditures are associated with cases that ultimately result in death. These patterns suggest that advanced age, more significant tumor burden, and higher treatment costs are associated with worse outcomes.

Categorical Variables

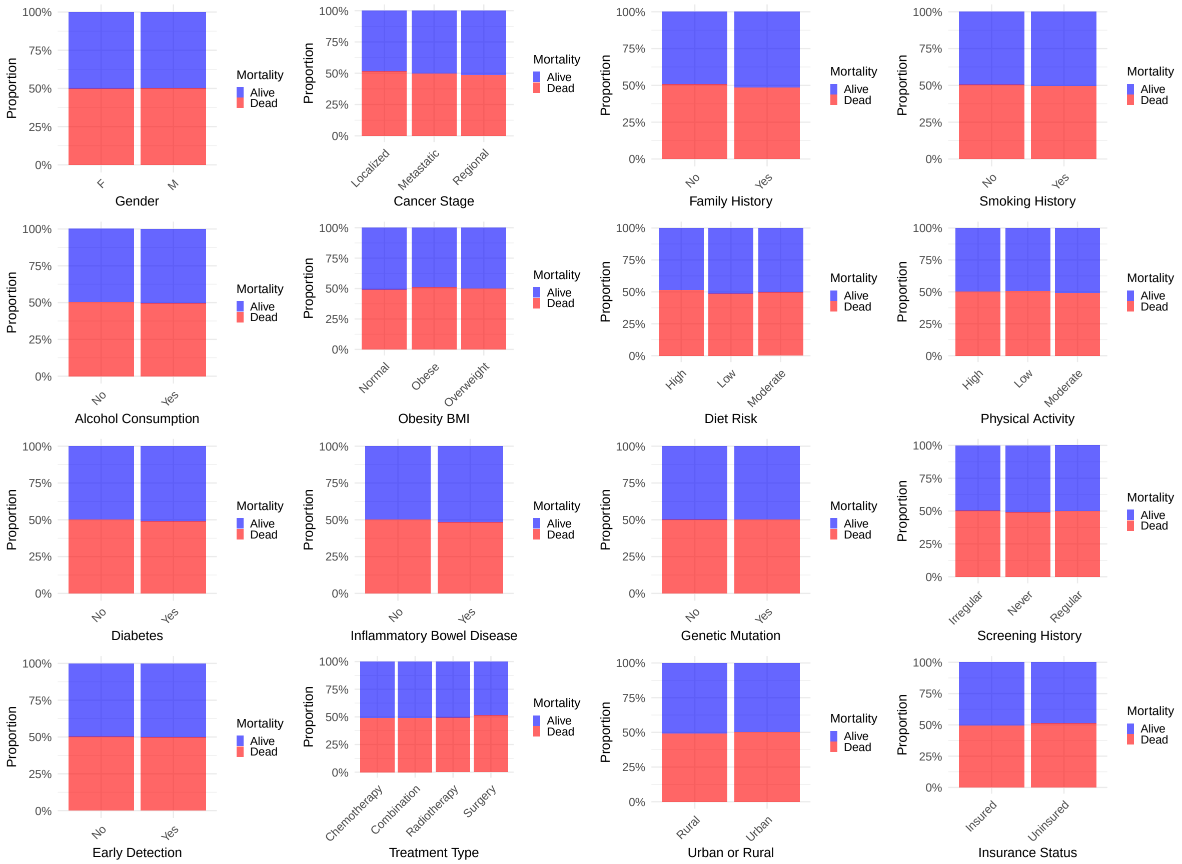

Figure 3: Proportion Bar plots for Categorical variables by mortality status

Interpretation of Findings

For the categorical variables, we plotted the proportions of patients' mortality status in each category (Figure 3) and observed some strong correlations. The cancer stage at diagnosis was a key factor, with patients with localized disease having a much higher mortality rate than patients with metastatic disease, and even patients with regional disease having a worse mortality rate than metastatic disease. Similarly, variables related to early intervention showed lower mortality in patients with a history of screening and in patients whose cancer was detected at an early stage compared with those who were not screened beforehand or who were not detected early. We also noted the influence of lifestyle and health risk factors, with categories such as smoking history, heavy alcohol consumption, obesity, poor diet and comorbidities tending to have higher mortality rates in the "yes" group than in the "no" group. In contrast, some factors showed little difference in mortality across categories, such as patient gender and family history of colorectal cancer, which did not show a significant difference in mortality status. Overall, advanced cancer and lack of early detection were associated with significantly higher mortality, while healthy behaviors and early cancer screening/detection had higher survival rates. These observed patterns helped us screen which variables may be significant predictors of mortality status, informing our modeling strategy.

Model Choice and Reasoning

Logistic Regression

Given the binary nature of the outcome (mortality: dead or alive), we chose a logistic regression model to quantify the association between the covariates and mortality. The model was applied to dichotomous responses and provided interpretable coefficients based on the odds ratio of mortality. EDA revealed a strong effect of cancer stage and early detection on the results, suggesting that a multivariate logistic approach would help analyze the model while controlling for other factors. Before fitting the model, we ensured that the key assumptions:

Assumption Check

Binary Response

The response variable is binary by definition, satisfying the requirement for logistic regression. We coded "Mortality" as a factor with two levels (Alive/Dead) and used the balanced dataset for modeling.

Independence

Each record corresponds to a unique patient, and we confirmed there were no duplicate entries. Thus, we can assume independence of observations.

Variance Structure

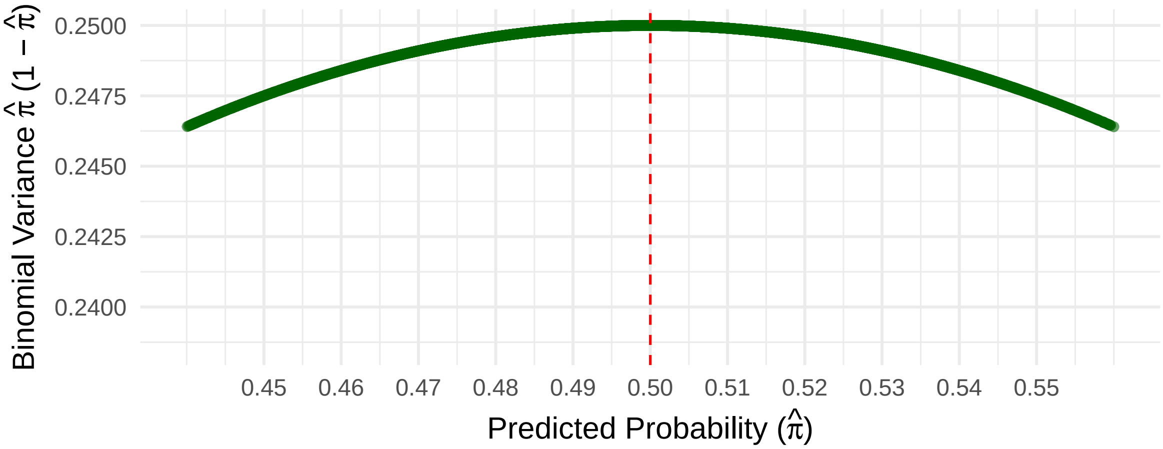

Figure 4: Variance peaks at predicted probability = 0.5.

In logistic regression, the variance of the residuals is not constant but is a function of the predicted probability $\hat{\pi}$. We plotted the predicted probabilities from the full model against the binomial variance (Figure 4). The plot showed that the variance is low at extreme predicted probabilities and peaks when $\hat{\pi}$ is almost equal to 0.5, forming the characteristic parabolic shape. The highest uncertainty occurs at about 50% of predicted risk of death, which is consistent with theory and suggests that the model's probability estimates are correct.

Linearity

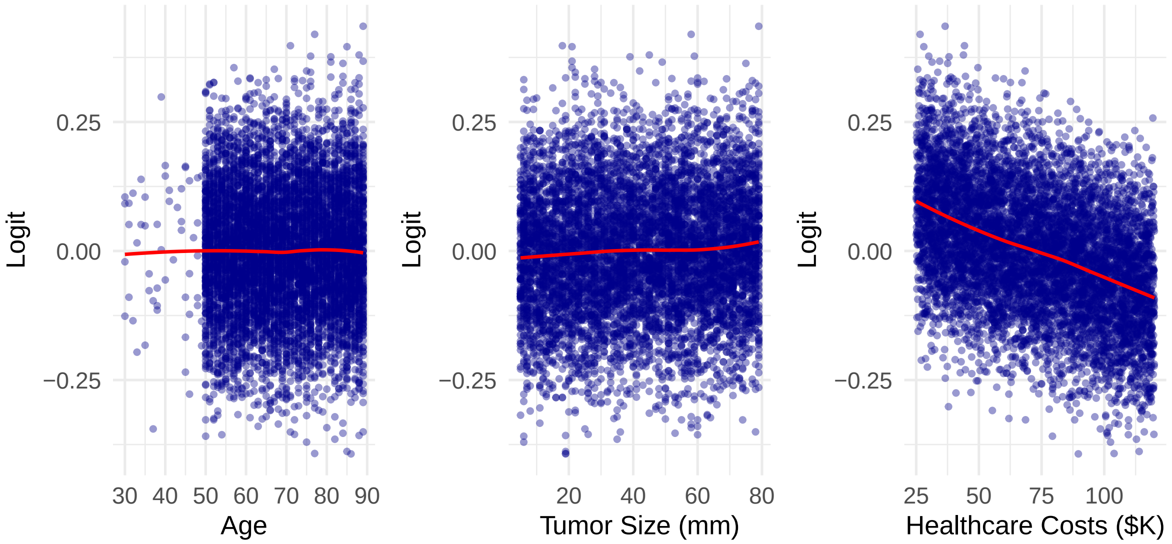

Figure 5: Linearity check − logit of mortality plotted against continuous variables

Figure 5 provides the linearity assumption by plotting the logit of predicted mortality against each continuous covariates of Age, Tumor Size, Healthcare Costs. The observed linear relations of red line confirm that this logistic regression assumption has been satisfied, and the plots for Age and Tumor

Size illustrates relatively constant trends, indicating minimal variation in the slope from linearity and it helps to prove the linear assumption for this study.

Feature Selection

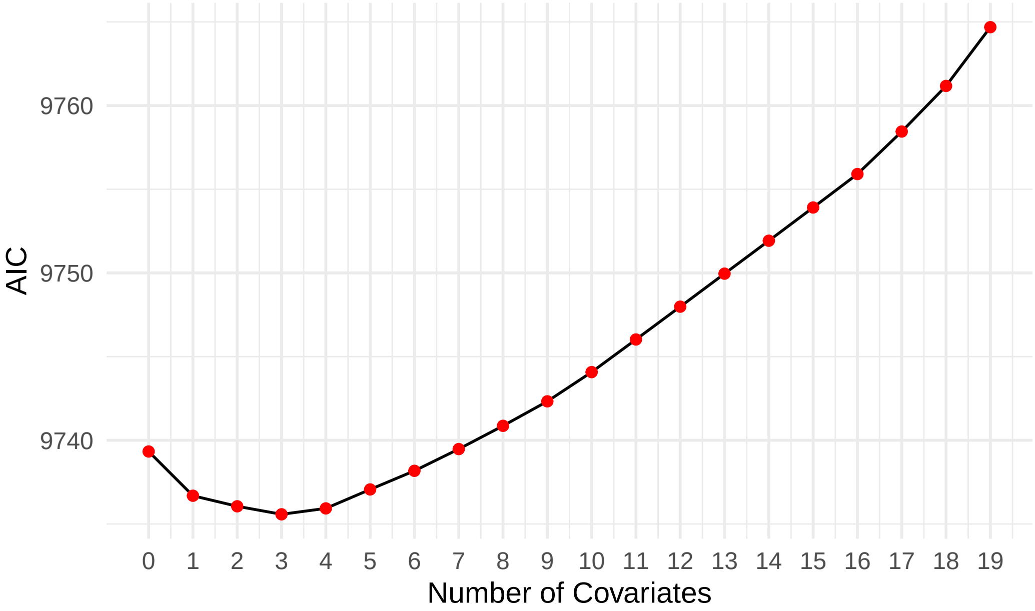

Figure 6: AIC vs Number of Covariates – demonstrating backward feature selection.

Model	Residual Deviance	AIC
Full Model	9710.7 (df = 6997)	9764.7
Backward-Selected Model	9725.6 (df = 7019)	9735.6

Table 1: Residual Deviance and AIC between the backward-selected and full models

The backward selection was performed for finding the optimal number of covariates for predicting mortality from colorectal cancer. This method iteratively removes the least significant variables from the full logistic regression model and selects the best model at each step based on the Akaike Information Criterion (AIC). After selecting the best model of this study, the selected model is evaluated against the full model to confirm the optimal balance between model interpretability (likelihood) and complexity (number of covariates).

Figure 6 illustrates how the AIC value changes during the backward selection process as variables are sequentially removed. Since the AIC preferred the smallest values, the lowest AIC was found in the plot when the number of covariates is 3, and also this process can lead to satisfy the principle of parsimony. Thus, the optimal model includes the following key predictors:

- **Cancer Stage**

- **Family History**
- **Healthcare Costs**

The AIC comparison between the full and backward selection models is summarized in Table 1. The full model, which includes all covariates, resulted in an AIC of 9764.7. In contrast, the optimal model has a lower AIC of 9735.6, indicating that the selected model provides a better balance between interpretability and complexity than the full model.

Statistical Analysis

To clarify whether including the interaction terms would significantly improve our model's predictive performance, we conducted a likelihood ratio test based on Wilks' theorem to compare the additive model against a model with interaction terms.

We defined the hypothesis as follows:

- Null hypothesis H_0: logistic regression of additive model is true model
- Alternative hypothesis H_a: logistic regression of interaction model is true model

To perform the hypothesis testing, we set the significance level at 5% (α = 0.05) and the result of test statistic and corresponding p-value are summarized below:

Test Statistic	**Degrees of Freedom**	**p-value**
$\chi^2 = 0.549$	1	0.4586

Table 2: Wilks' likelihood ratio test comparison for Additive and Interaction

In table 2, the observed p-value is 0.4586 and it is greater than our pre-specified significance level of α = 0.05. Thus, we fail to reject the null hypothesis and conclude that including interaction terms does not significantly improve model fit. Consequently, simpler additive model of this study ensures interpretability and prevents overfitting by the principle of parsimony.

Predictor	**Estimate**	**Std. Error**	**z value**	**Pr($>|z|$)**	**Significance**
(Intercept)	0.2256713	0.0804475	2.805	0.00503	**
Cancer_StageMetastatic	-0.0707315	0.0648033	-1.091	0.27506	
Cancer_StageRegional	-0.1151017	0.0537461	-2.142	0.03223	*
Family_HistoryYes	-0.0817314	0.0520083	-1.572	0.11607	
Healthcare_Costs	-0.0018622	0.0008735	-2.132	0.03301	*
Early_DetectionYes	-0.0081611	0.0487565	-0.167	0.86707	

Table 3: Regression output with significance levels: * p < 0.05, ** p < 0.01

Results Interpretation

Since the primary goal of this report was to evaluate how healthcare expenditures and early detection of colorectal cancer associated, so the final model included the covariate “Early Detection” with selected variables from the backward selection. To clarify whether including an interaction term between **Early Detection** and **Healthcare Costs** improves the model fit, additive model and one with the interaction term with two covariates model were conducted the likelihood ratio test (LRT) and the results in comparison of goodness of fit between an additive model and a model with an interaction term between the two covariates using ANOVA were summarized in Table 3.

- **Cancer Stage (Regional):** Since the p-value of cancer stage at regional is statistically significant, thus we conclude that patients diagnosed at the regional stage have significantly lower odds of cancer mortality compared to the localized stage.
- **Healthcare Cost:** The observed p-value is 0.03301, indicating inverse relationship with mortality and higher spending for patient may correspond to more effective treatments and better outcomes.
- **Family History and Early Detection:** They are not statistically significant at the 5% level, but Early Detection was retained for the objectives of this study and its practical and theoretical role in influencing the early stage of cancer.

Conclusion

Main Findings

This analysis identified colorectal cancer stage at diagnosis and patient health care costs as significant predictors of mortality from colorectal cancer. Specifically, mortality was much higher for advanced cancers (especially metastatic disease) compared to localized cancers, consistent with evidence that stage at diagnosis is the most important prognostic factor for colorectal cancer (Du Cancer, 2024). Higher healthcare expenditure was also significantly associated with mortality risk, which may reflect the intensive treatment needs and complications of advanced disease (Balkhi et al., 2023). In contrast, early detection status was not an independent predictor of mortality after controlling for cancer stage and other factors. This finding suggests that the benefits of early detection are mediated primarily through its impact on the diagnostic stage rather than providing an additional survival advantage (McPhail et al., 2015). These results emphasize the importance of early diagnosis of colorectal cancer in improving survival. Thus, the findings support the management of colorectal cancer patients through effective screening programs and allocating adequate healthcare resources to improve patient outcomes, emphasizing the importance of early detection.

Limitations

- This observational study limits the ability to establish causal relationships between predictors and mortality outcomes.
- Unmeasured confounders may also have affected the observed associations.
- The analysis lacks variables that represent systemic and contextual factors that may influence outcomes, such as regional differences in health care services or socioeconomic status.
- The variable “early detection” is defined very broadly (yes/no indicator of early detection), making its meaning somewhat ambiguous.

Potential Further research

Future studies could use more in-depth analyses to validate and extend these findings and examine the impact of interventions. For example, survival analyses (e.g., using Cox proportional risk models) could incorporate time-to-event data to determine in greater detail when, not just if, death occurs (Collett, 2015).

Appendix

- Full regression output
- Extra plots or tables not essential to the main body
- Model selection steps

References

Balkhi, B., Alghamdi, A., Alqahtani, S., Najjar, M. A., Harbi, A. A., & Traiki, T. B. (2023). Colorectal cancer-related resource utilization and healthcare costs in Saudi Arabia. *Saudi Pharmaceutical Journal, 31*(11), 101822. https://doi.org/10.1016/j.jsps.2023.101822

Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC eBooks. https://doi.org/10.1201/b18041

Canadian Cancer Society. (2024, May 1). *Colorectal cancer statistics*. https://cancer.ca/en/cancer-information/cancer-types/colorectal/statistics

McPhail, S., Johnson, S., Greenberg, D., Peake, M., & Rous, B. (2015). Stage at diagnosis and early mortality from cancer in England. *British Journal of Cancer, 112*(S1), S108–S115. https://doi.org/10.1038/bjc.2015.491