# 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***

### Continuous Variables

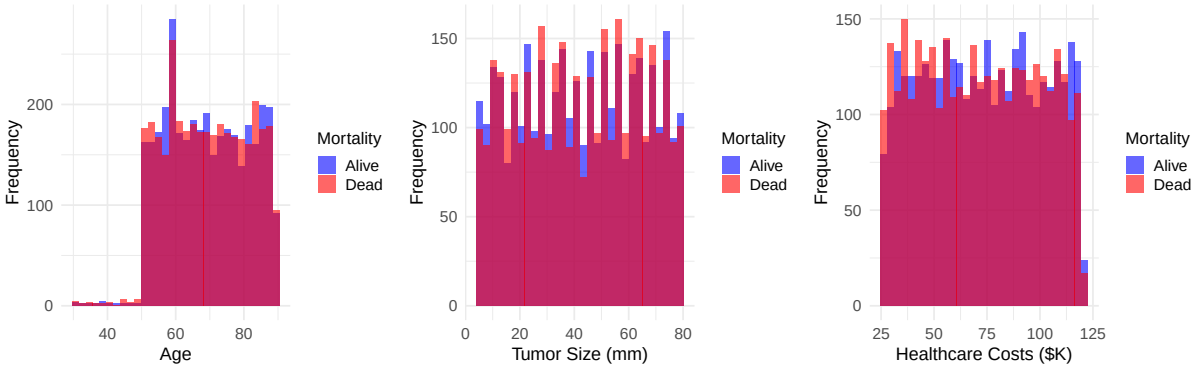


***Figure 2: Distribution of continuous variables by mortality status***

## Categorical Variables

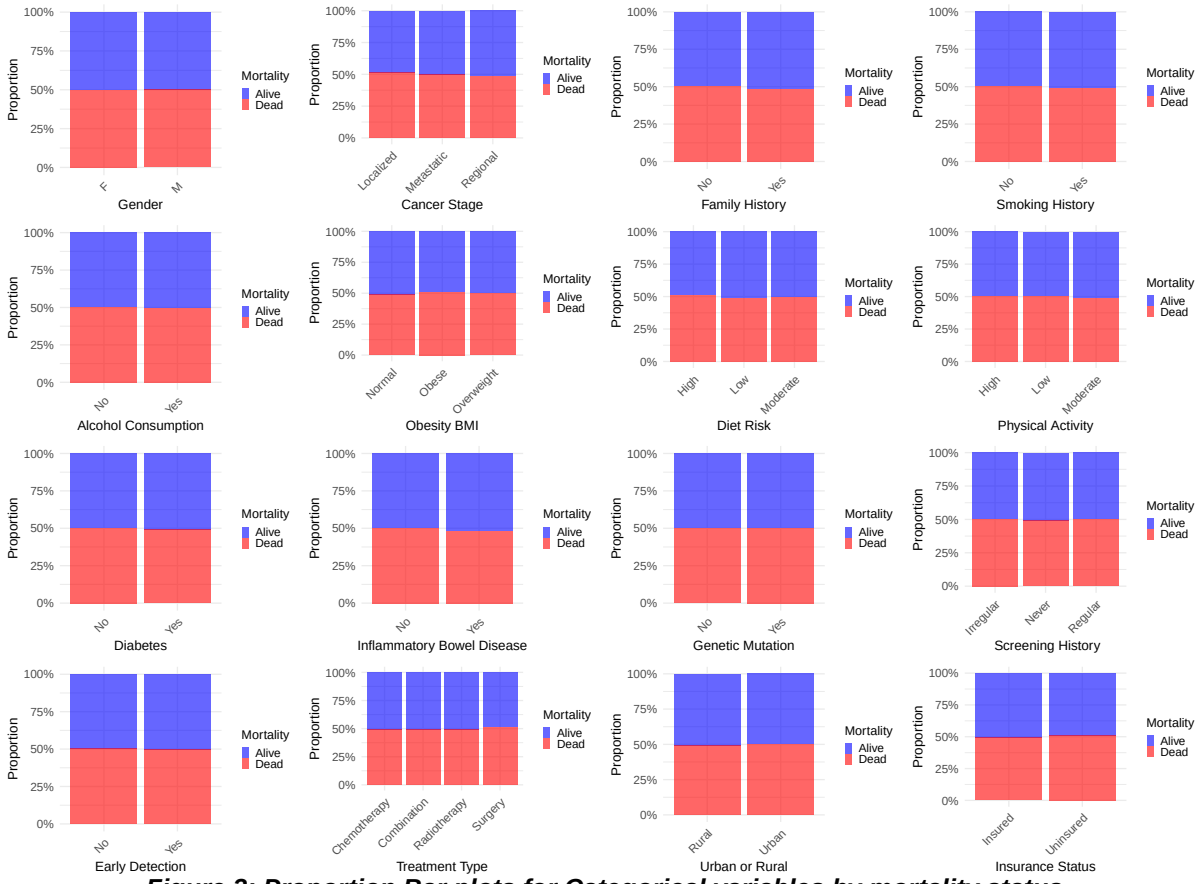


***Figure 3: Proportion Bar plots for Categorical variables by mortality status***

## Interpretation of Findings

Pattern, trends, suggested operations

## Model Choice and Reasoning

## Logistic Regression

explain why choose this model based on EDA and Data description

## Assumption Check

1. **Binary Response**

   Based on *Figure 1* , the response variable is binary

2. **Independence**

   No duplicate rows, independence hold

3. **Variance Structure**

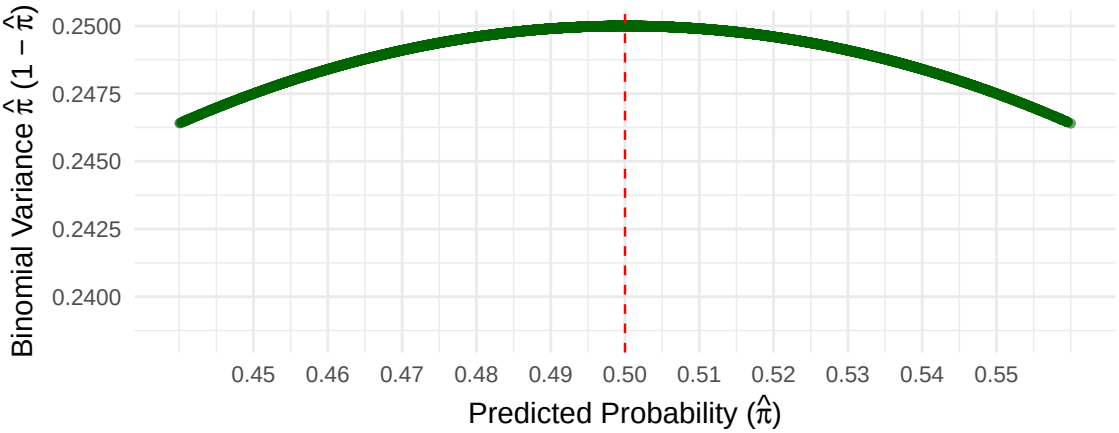


***Figure 4: Variance peaks at predicted probability = 0.5.***

4. **Linearity**

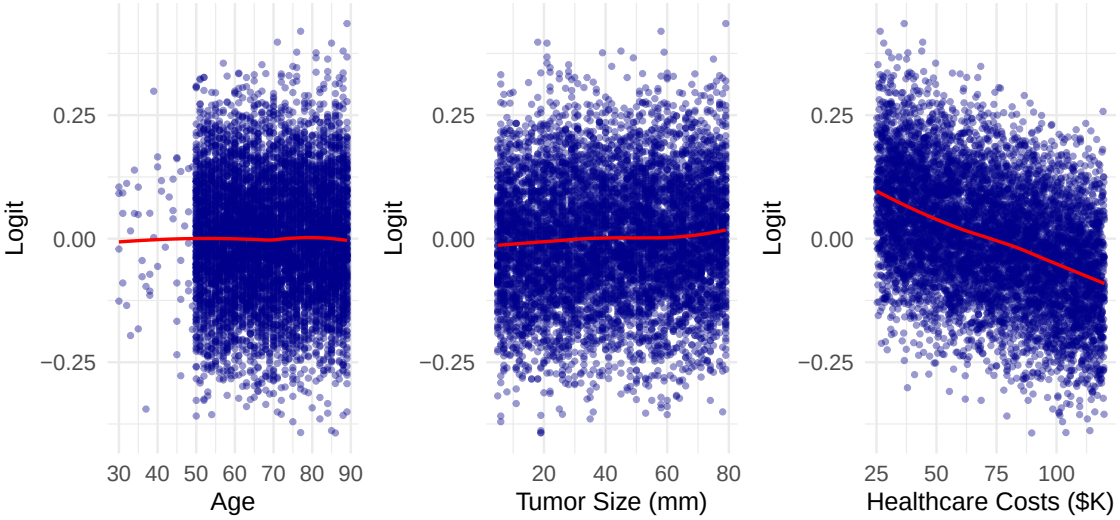


***Figure 5: Linearity check − logit of mortality plotted against continuous variables***

## Feature Selection

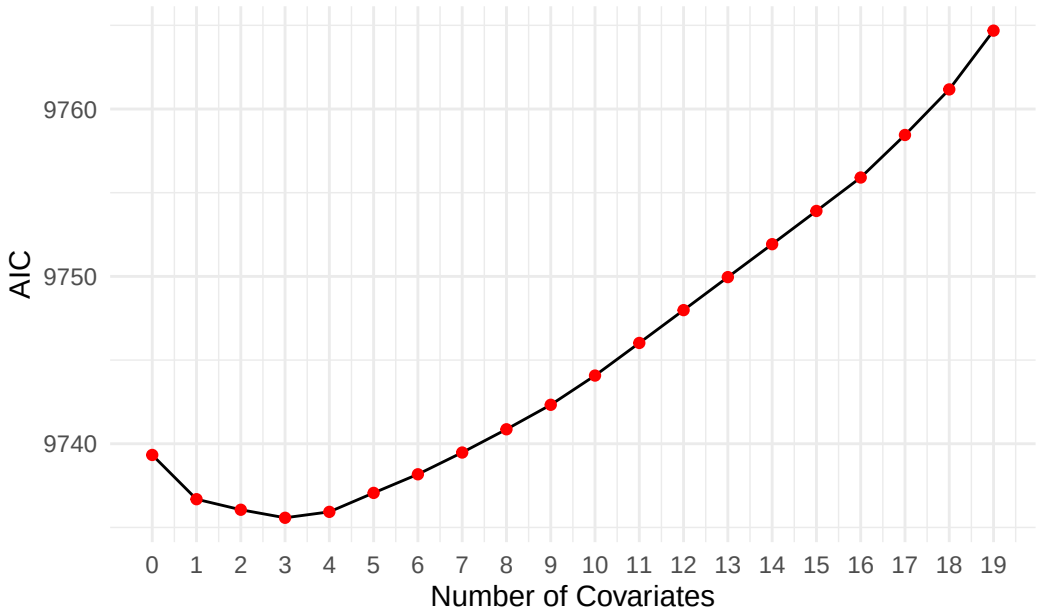


***Figure 6: AIC vs Number of Covariates − demonstrating backward feature selection.***

| 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

…

### Statistical Analysis

| 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

The result suggests the interaction term does not significantly improve model fit.

…

…

### Results Interpretation

inference results

| 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

## 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

## Reference

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. In *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/b18041

Du Cancer, C. C. S. /. S. C. (2024, May 1). *Colorectal cancer statistics*. Canadian Cancer Society. 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.2015.49