Associations of Healthcare Costs and Early Detection with Colorectal Cancer Mortality in Canada

Yulin Yuan, Ruohan Sun, Tony Lee

April 4, 2025

Introduction

Study Objective

Clearly define the research questions or objectives.

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

Variable	**Description**	**Type**
Mortality Status	Status of patients' mortality (Yes/No)	Categorical, nominal

Explanatory Variables

Variable	**Description**	**Type**
Alcohol Consumption	Status of patient's alcohol consumption (Yes/No)	Categorical, nominal
Age	Patient's age in years	Numerical, continuous
Cancer Stage	Diagnosis stage of cancer (Localized, Regional, Metastatic)	Categorical, ordinal
Diabetes	Status of patient's diabetes (Yes/No)	Categorical, nominal
Diet Risk	Level of dietary risk (Low, Moderate, High)	Categorical, ordinal
Early Detection	Detection of colorectal cancer at an early stage (Yes/No)	Categorical, nominal
Family History	Presence of family history of colorectal cancer (Yes/No)	Categorical, nominal
Genetic Mutation	Presence of genetic mutations for colorectal cancer (Yes/No)	Categorical, nominal
Gender	Gender of the patient (Male/Female)	Categorical, nominal
Healthcare Cost	Estimated healthcare expenditure per patient (1,000 units in $)	Numerical, continuous
Inflammatory Bowel Disease	Status of inflammatory bowel disease (Yes/No)	Categorical, nominal
Insurance Status	Health insurance coverage (Insured/Uninsured)	Categorical, nominal

Obesity BMI	BMI classification (Normal, Overweight, Obese)	Categorical, ordinal
Physical Activity	Level of physical activity (Low, Moderate, High)	Categorical, ordinal
Screening History	Cancer screening history (Regular, Irregular, Never)	Categorical, ordinal
Smoking History	Patient's smoking history (Yes/No)	Categorical, nominal
Treatment Type	Type of treatment received (Surgery, Chemotherapy, Radiotherapy, Combination)	Categorical, ordinal
Tumor Size	Colorectal tumor size in millimeters (mm)	Numerical, continuous
Urban or Rural	Patient's area of residence (Urban/Rural)	Categorical, nominal

Motivation

- Why is this question or data important/interesting?
- Need to explain why we remove country-level variable and Patient_ID, reasonable is fine.

Analysis

Exploratory Data Analysis (EDA)

Balance of Response Variable

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

Continuous Variables

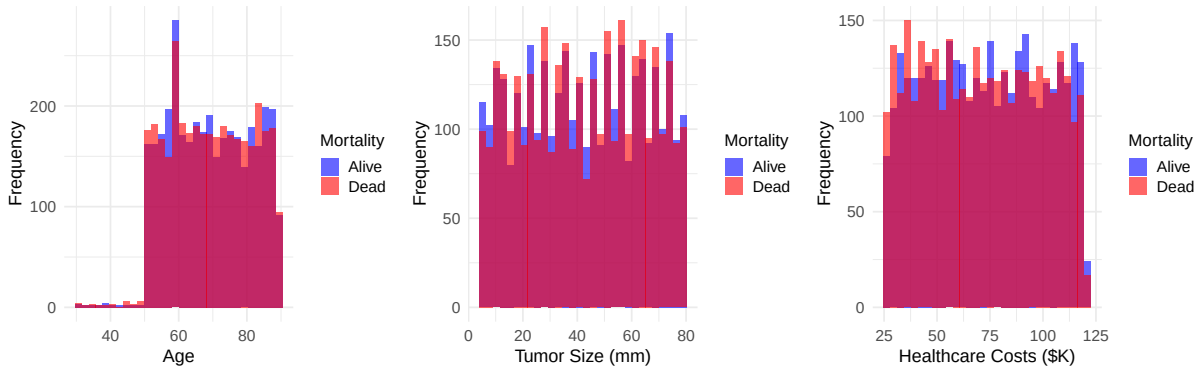

Figure 2: Distribution of continuous variables by mortality status

Categorical Variables

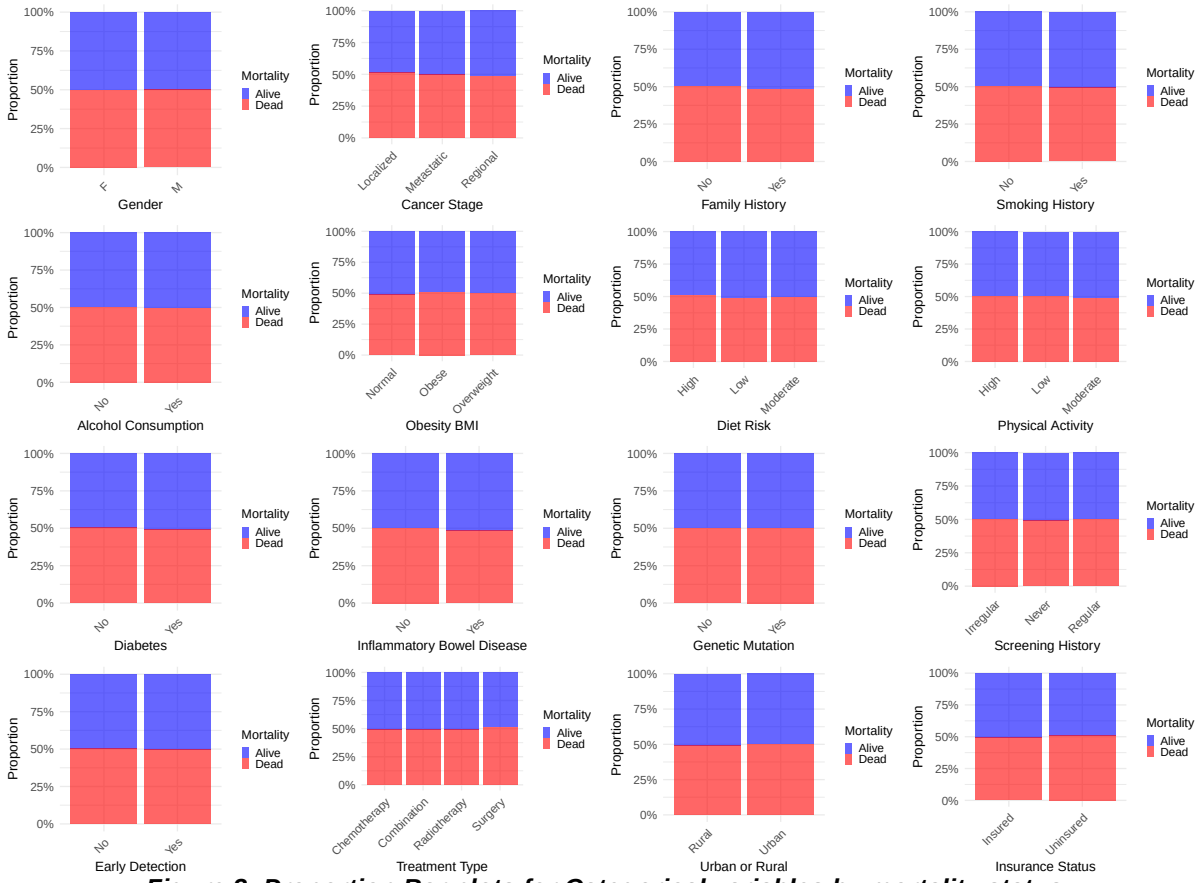

Figure 3: Proportion Bar plots for Categorical variables by mortality status

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**

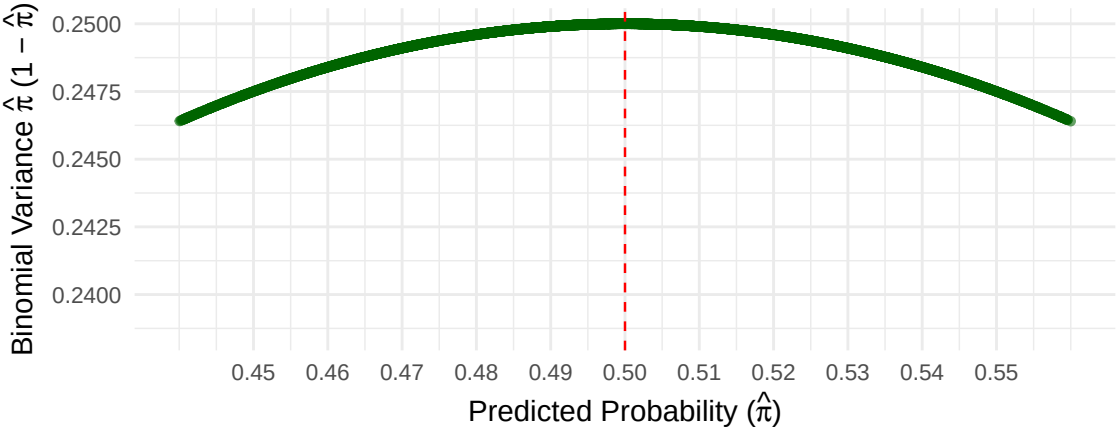

Figure 4: Variance peaks at predicted probability = 0.5.

4. **Linearity**

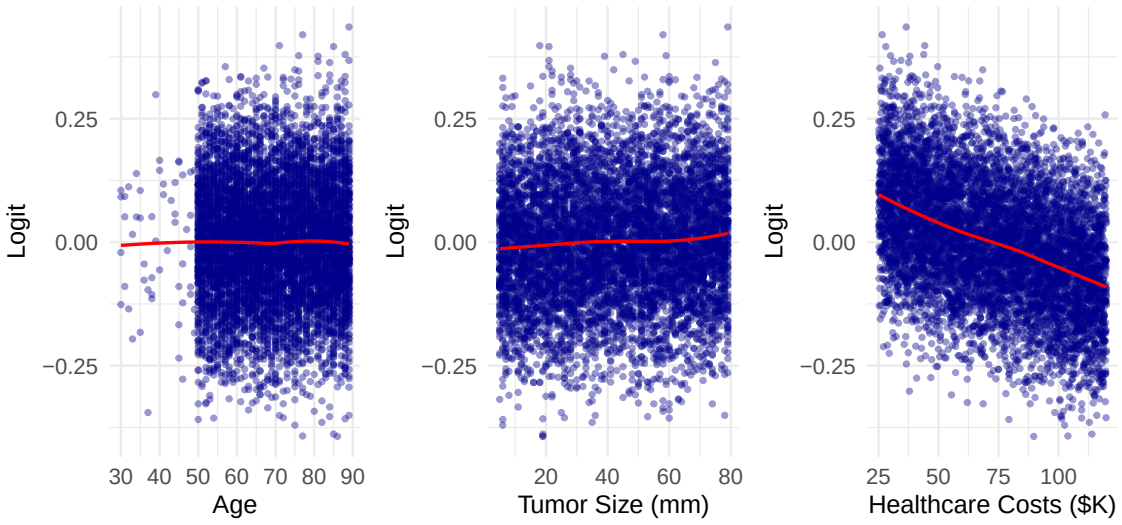

Figure 5: Linearity check − logit of mortality plotted against continuous variables

Feature Selection

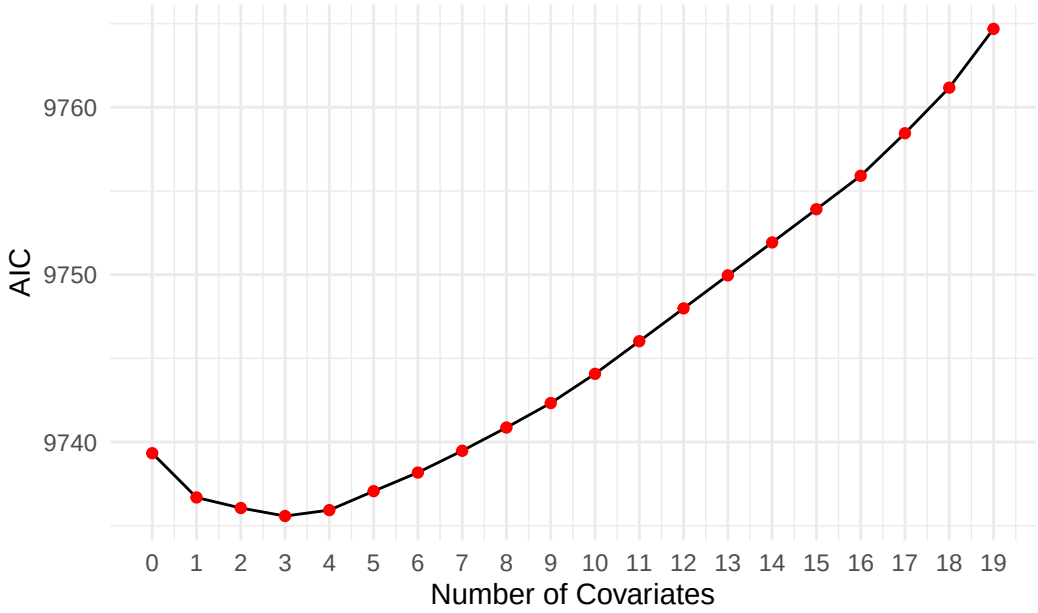

Figure 6: AIC vs Number of Covariates − demonstrating backward feature selection.

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.

...

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

inference results

Conclusion

Main Findings

Interpreting result in real-world context, careful about causality

Limitations

- Discuss possible sources of bias, limitations in data, model assumptions
- Suggest improvements or next steps

Potential Further research

Mention anything interesting you found that doesn't fit elsewhere

Appendix

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