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

Yulin Yuan, Ruohan Sun, Tony Lee

Introduction

Study Objective

Clearly define the research questions or objectives.

Dataset Overview

Describe:

- Where the data comes from
- What each variable represents
- How the data was collected (if known)

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

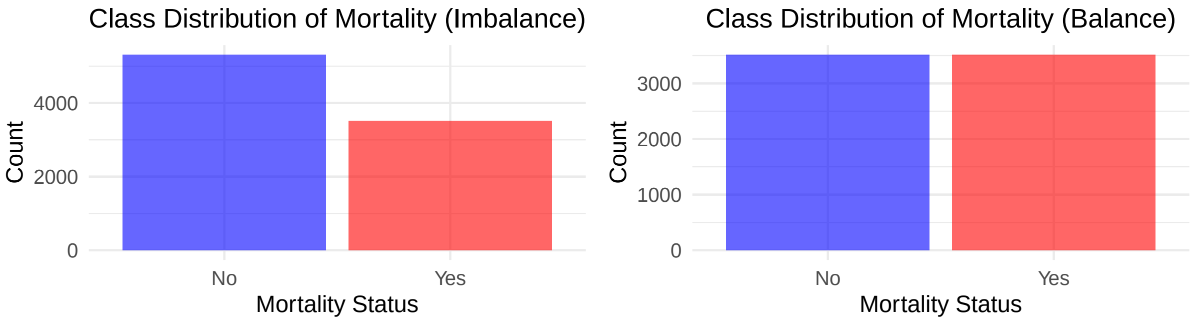

Figure 1: Class Distribution of the Mortality Variable Before and After Balancing

Continuous Variables

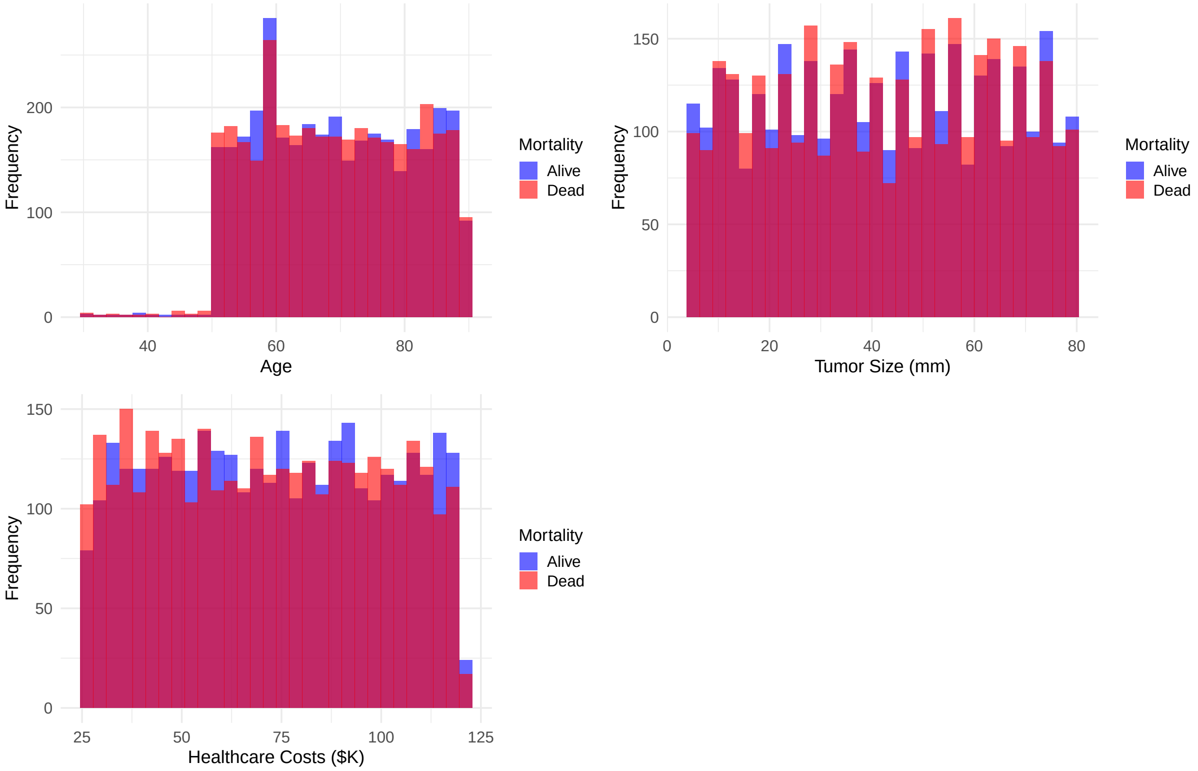

Figure 2: Distribution of continuous variables by mortality status

Categorical Variables

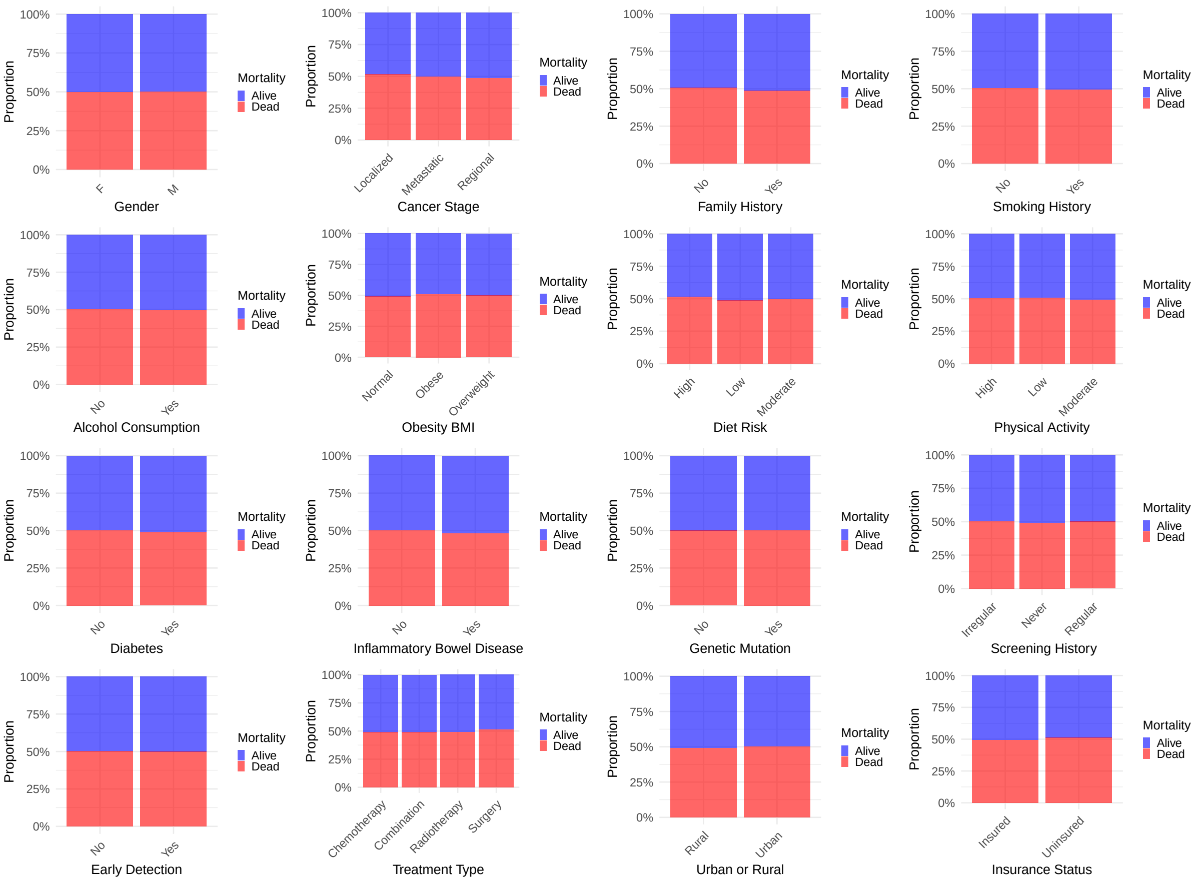

Figure 3: Proportion Bar plots for Categorical variables by mortality status

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**

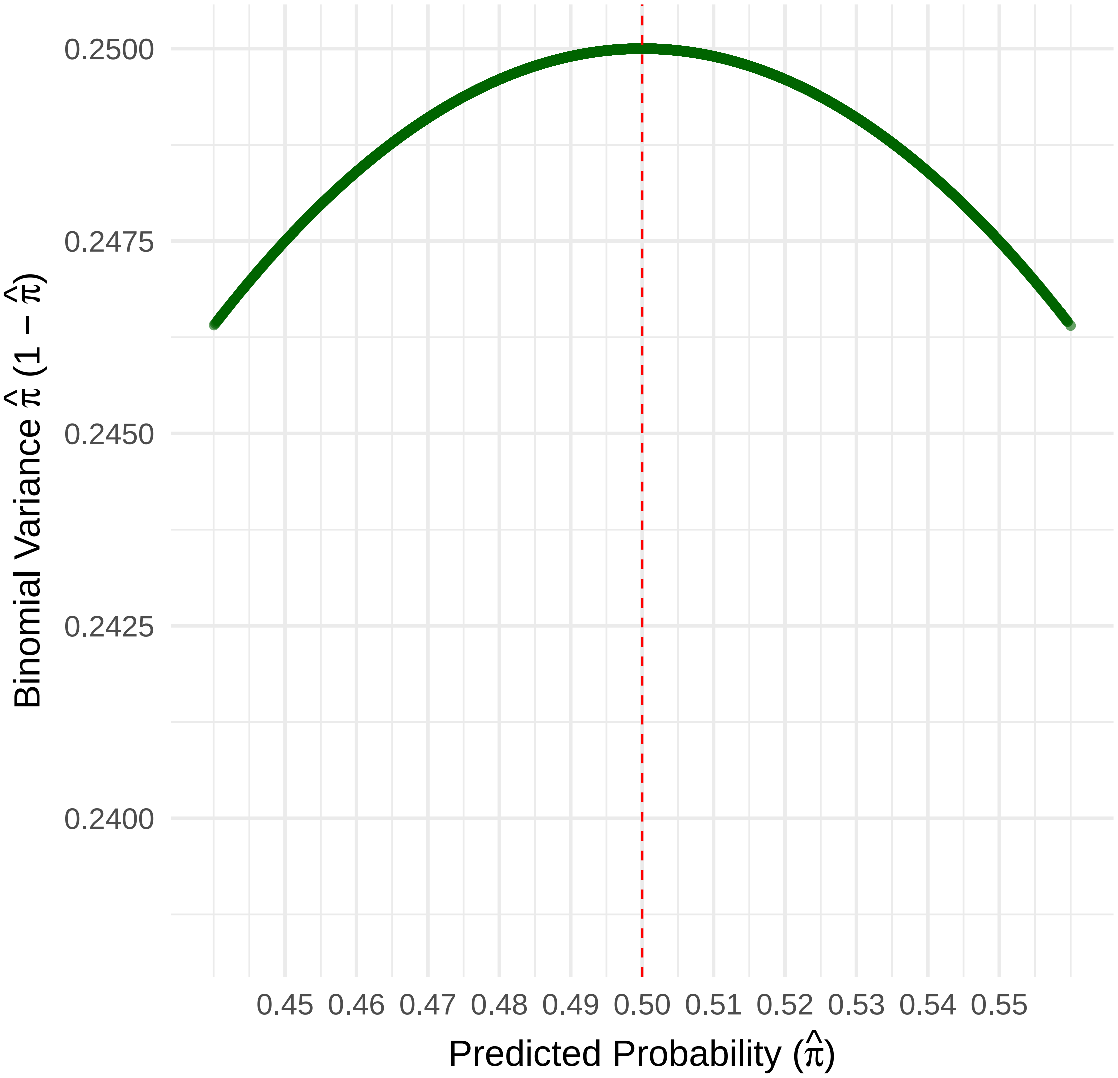

Figure 4: Variance peaks at predicted probability = 0.5.

4. **Linearity**

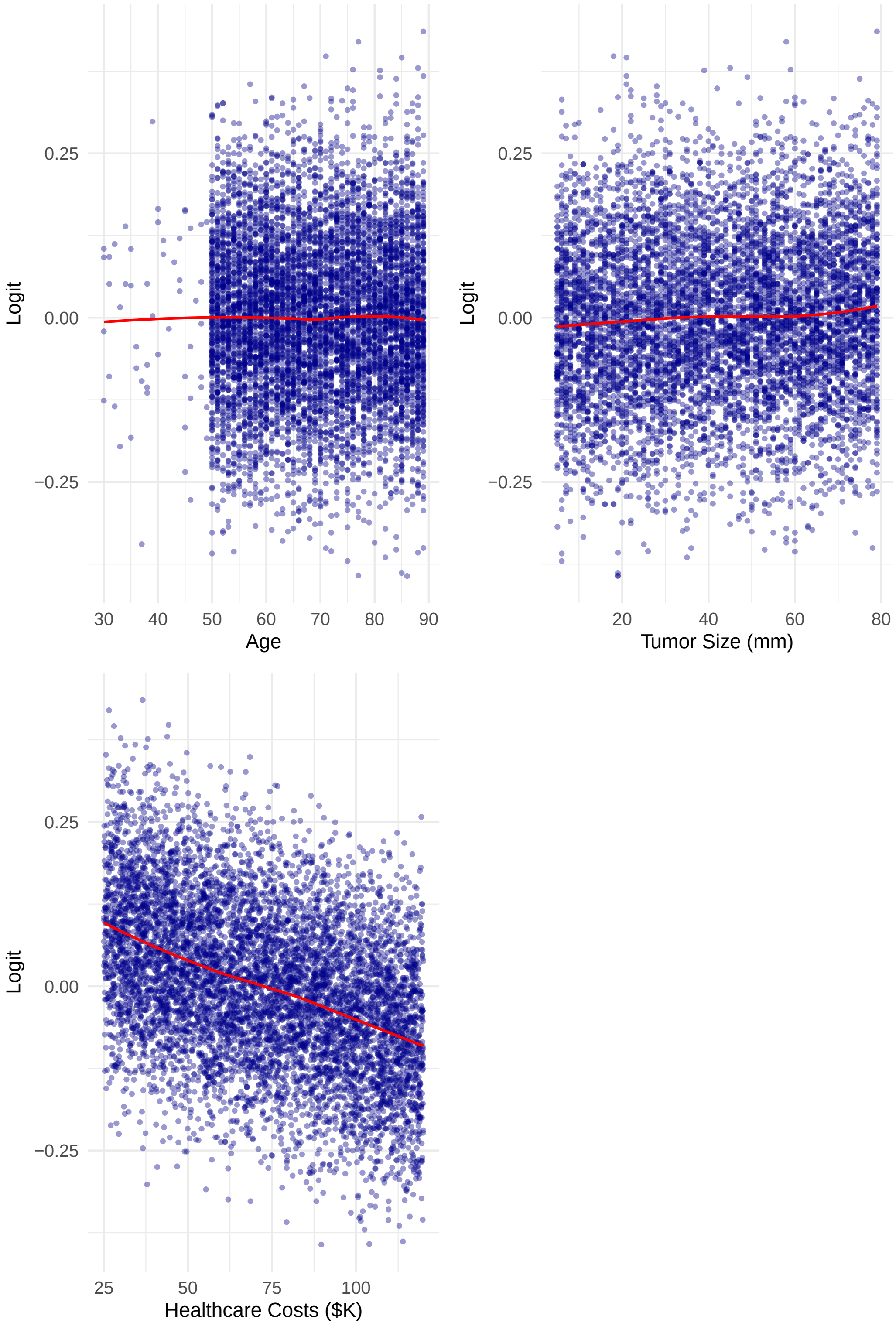

Figure 5: Linearity check − logit of mortality plotted against continuous variables

Feature Selection

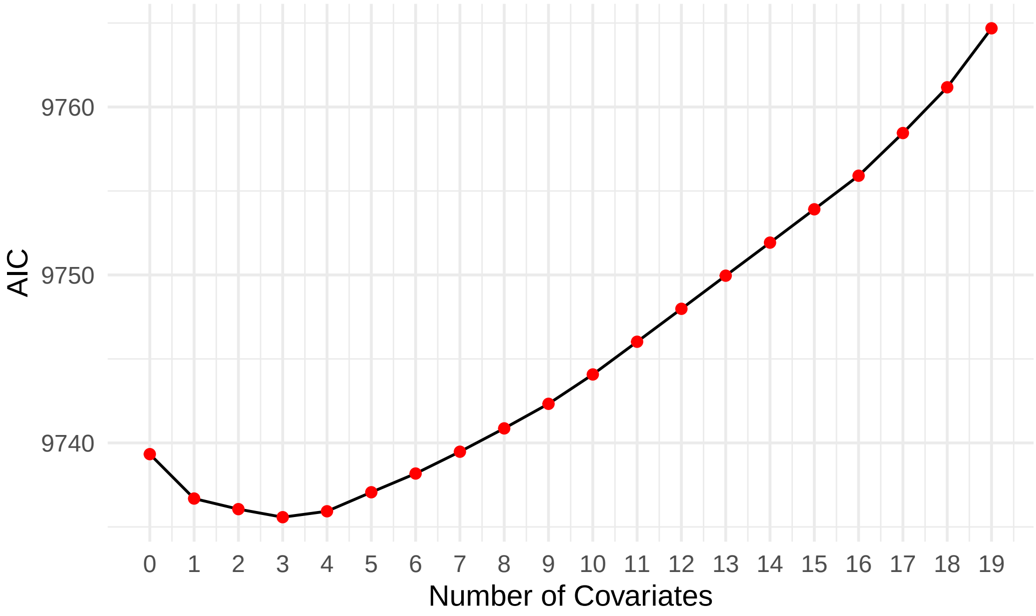

Figure 6: AIC vs Number of Covariates – demonstrating backward feature selection.

```
Call:
glm(formula = Mortality ~ Cancer_Stage + Family_History + Healthcare_Costs,
    family = binomial, data = balanced_data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             0.2209489  0.0753360   2.933  0.00336 **
Cancer_StageMetastatic -0.0707645  0.0648029  -1.092  0.27484
Cancer_StageRegional   -0.1152779  0.0537357  -2.145  0.03193 *
Family_HistoryYes      -0.0818704  0.0520016  -1.574  0.11540
Healthcare_Costs       -0.0018629  0.0008734  -2.133  0.03294 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9737.3  on 7023  degrees of freedom
Residual deviance: 9725.6  on 7019  degrees of freedom
AIC: 9735.6

Number of Fisher Scoring iterations: 3
```

Statistical Analysis

```

Call:
glm(formula = Mortality ~ Cancer_Stage + Family_History + Healthcare_Costs +
    Early_Detection, family = binomial, data = balanced_data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9737.3  on 7023  degrees of freedom
Residual deviance: 9725.5  on 7018  degrees of freedom
AIC: 9737.5

Number of Fisher Scoring iterations: 3
```

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