Predicting Thyroid Cancer Recurrence

Joyce Gill, Matthew Billings-Chiu

2025-05-02

Contents

0.1 Introduction

Cancer is the unregulated proliferation of cells which consume dangerous and disproportionately high amounts of energy from the body. These cells can shunt blood flow to maintain their high metabolic cost and increasing size (potentially blocking organ function), which becomes deadly, especially if it metastasizes (spreading elsewhere in the body). In the United States, cancer is the leading cause of death for people under 65 years old. We plan to study thyroid cancer specifically, which is projected to have 44,000 new cases this year, and is unique in that (although statistically insignificant) its death rate has trended upwards in recent years (whereas many have had a statistically significant decrease).1 Thyroid cancer occurs in the thyroid, an endocrine system gland located at the base of the neck that regulates heart rate, blood pressure, body temperature and weight hormonally.

Luckily, thyroid cancer has treatment options, including surgical removal and Radioactive Iodine therapy (RAI) that targets cancerous thyroid cells by exposing them to radioactive iodine (which is primarily taken up by the thyroid). However, treatment response can vary, and monitoring how well a patient responds is critical. We believe that a poor response may increase the likelihood of recurrence. As such, understanding factors such as treatment response on recurrence is highly valuable for deciding a patient's next steps.

This study investigates the following research question: Holding other relevant factors constant (age, gender, prior radiotherapy, and clinical risk classification), how does initial treatment response predict thyroid cancer recurrence? Our findings indicate that treatment response is a statistically significant predictor of recurrence when controlling for age, gender, radiotherapy history, and risk classification. We also explore the relationships among these explanatory variables for further statistical analysis.

0.2 Methods

The dataset for this study was sourced from Kaggle and originates from a published article by Hamadan University in the European Archives of Oto-Rhino-Laryngology. It contains data on 383 thyroid cancer patients, each of whom was followed for at least 10 years, with records spanning over a 15-year period. The dataset includes information related to patient demographics, treatment history, and clinical outcomes.

To answer our research question, we examined the following variables:

- Recurrence (Binary Categorical Dependent Variable): Whether or not cancer recurred.
- Age (Quantitative Explanatory Variable): Patient's age in years.
- Gender (Binary Categorical Explanatory Variable): Male or female.
- Radiotherapy (Binary Categorical Explanatory Variable): History of prior radiotherapy (yes or no).
- Risk (3 Level Categorical Explanatory Variable): Cancer risk classification (low, medium, high).
- Response (4 Level Categorical Explanatory Variable): Initial treatment response (excellent, indeterminate, structural incomplete, biochemical incomplete).

Since our primary predictors are categorical, the assumption of linearity in the logit is satisfied by the nature of categorical predictors. Additionally, the assumptions of independence and randomness were reasonably met based on the data collection methodology.

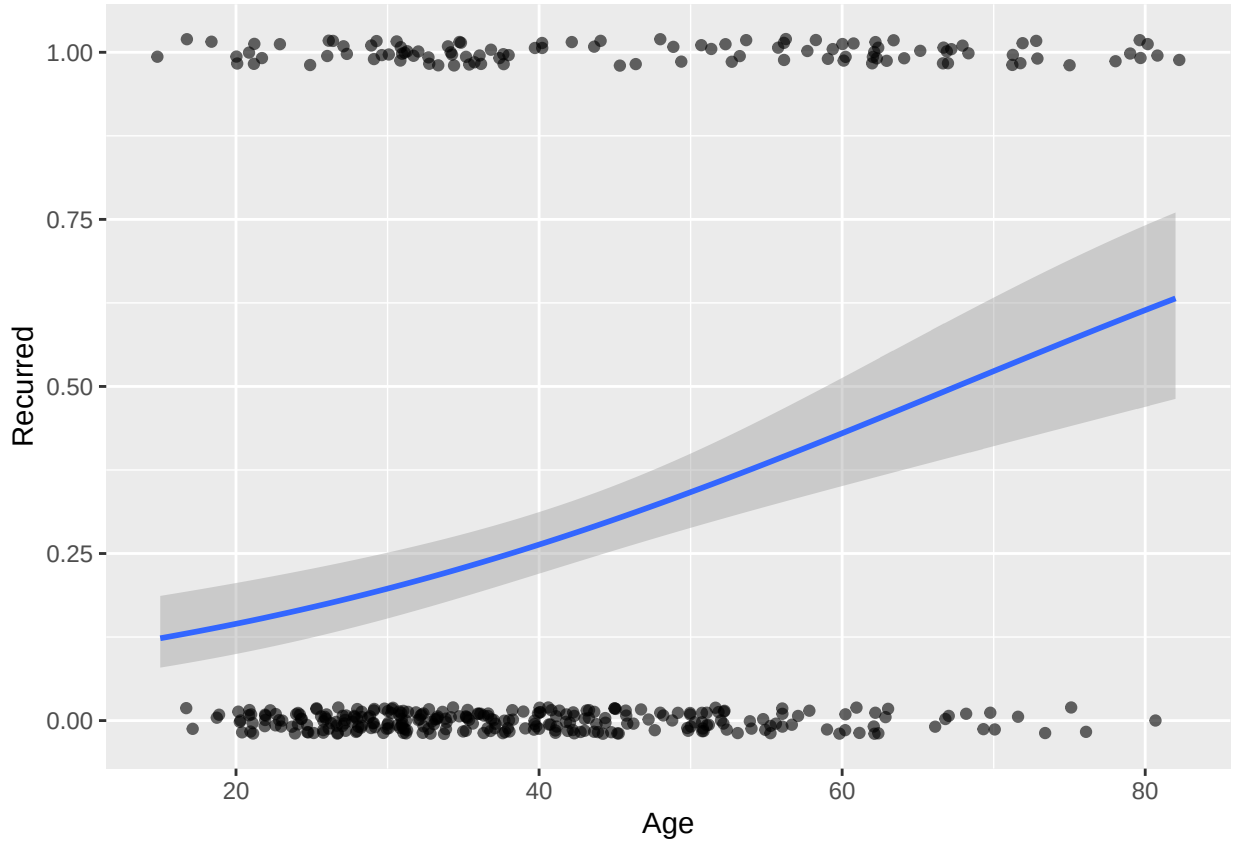
For further analysis, we recoded the Recurrence variable as a binary indicator (0 = no recurrence, 1 = recurrence) and collapsed Risk into a binary variable (high vs. not high). We then fit a multiple logistic regression model using recurrence as the response variable and the following as predictors: treatment response, age, gender, radiotherapy history, and binary risk classification. The fitted model is structured as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = 24.05538$$
$$- 4.69692\,(\text{ResponseExcellent})$$
$$- 1.66295\,(\text{ResponseIndeterminate})$$
$$+ 4.00775\,(\text{ResponseStructuralIncomplete})$$
$$+ 0.03017\,(\text{Age})$$
$$+ 1.04256\,(\text{GenderM})$$
$$- 10.63361\,(\text{Hx.RadiotherapyYes})$$
$$- 26.05770\,(\text{risk\_binaryNotHigh})$$

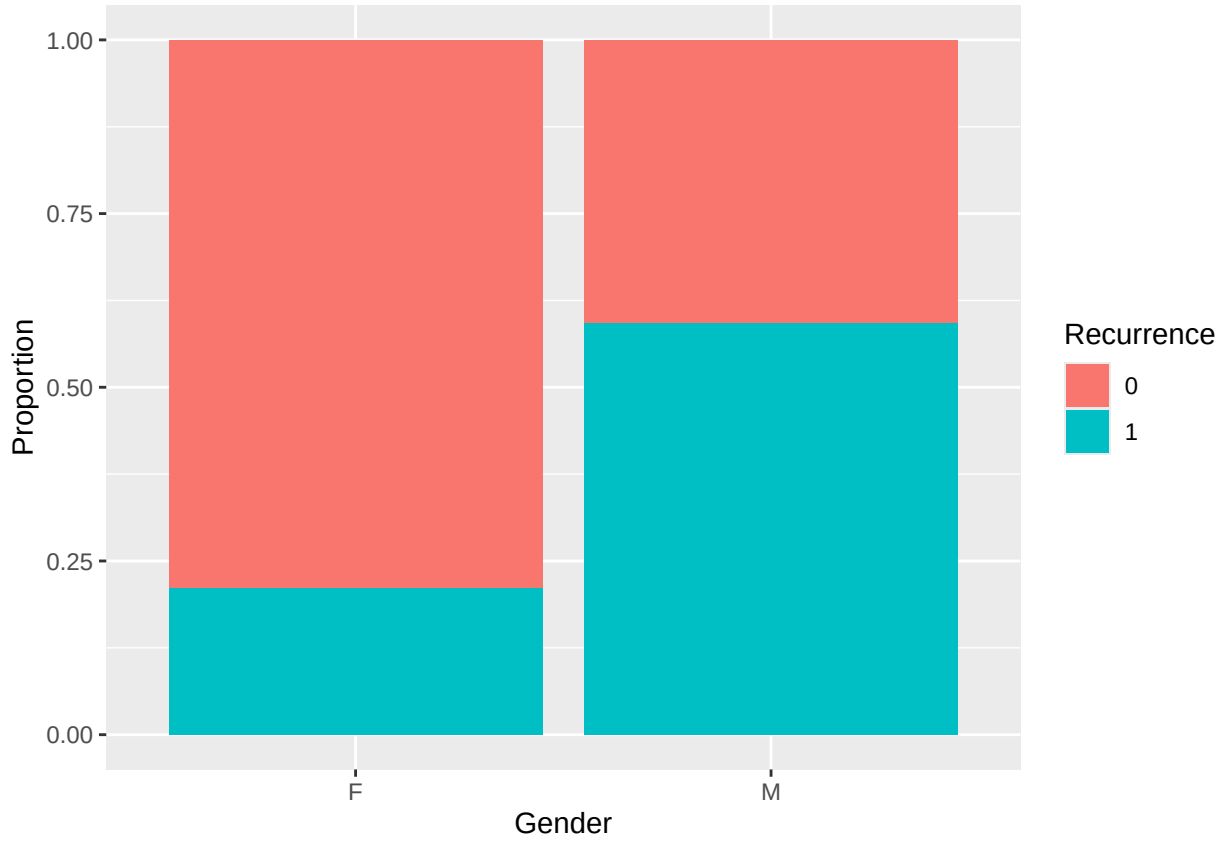
0.3 Analysis and Results

```
ggplot(data, aes(x = Age, y = Recurred)) +
  geom_jitter(height = 0.02, alpha = 0.6) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"))
```

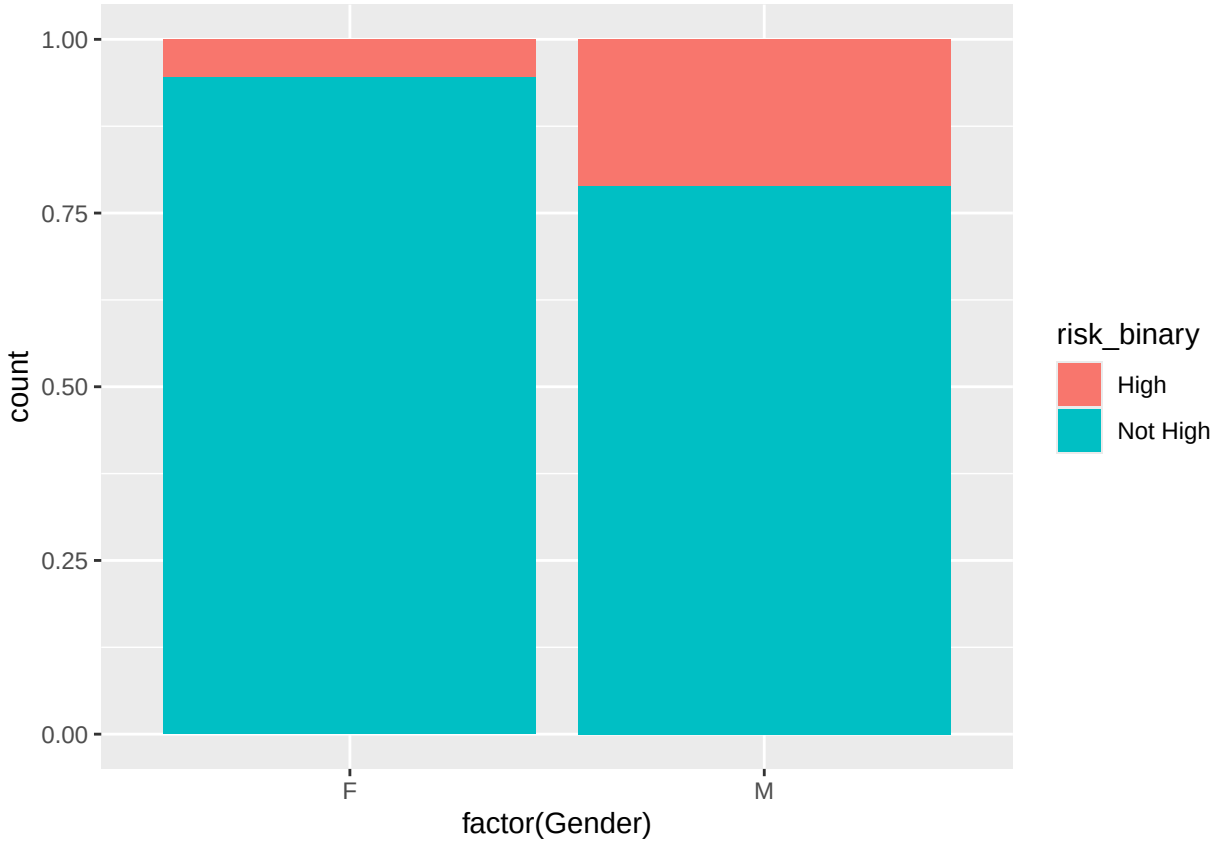
```
## `geom_smooth()` using formula = 'y ~ x'
```



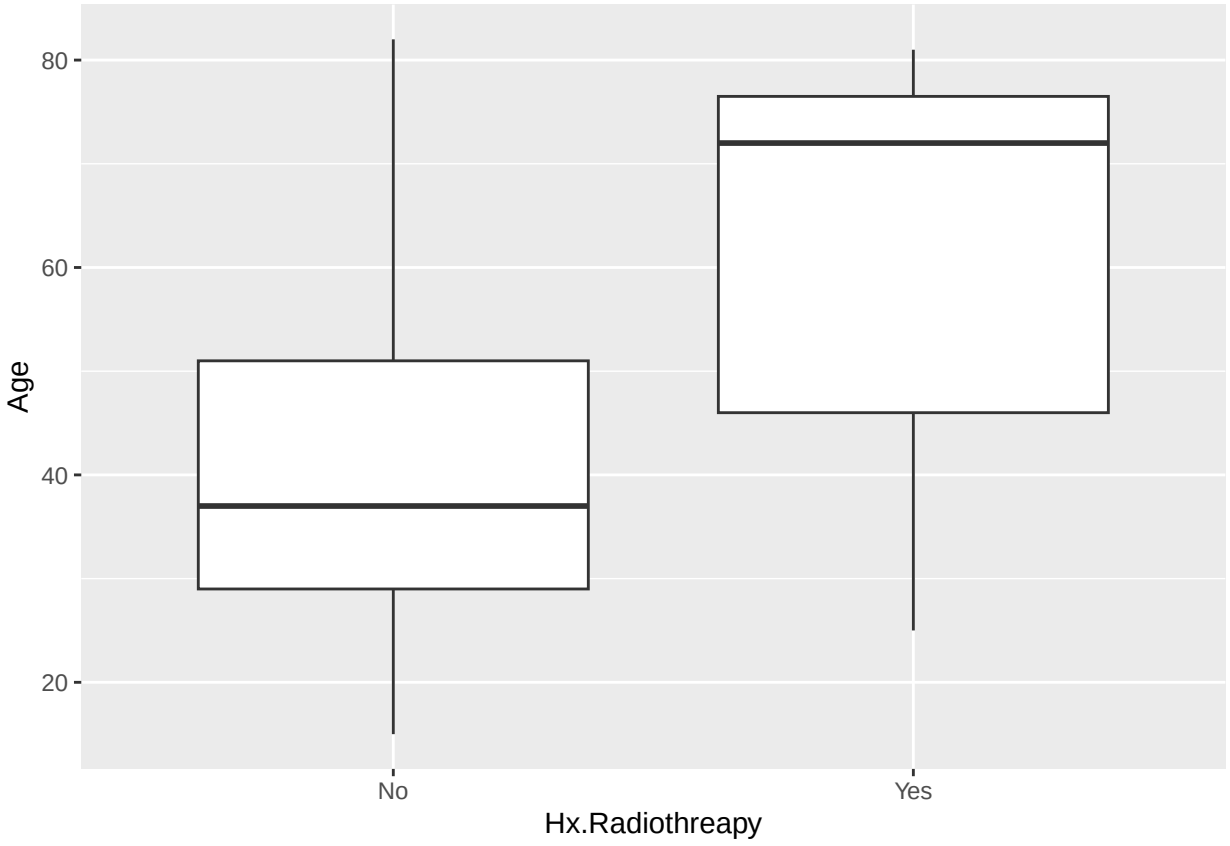
```
ggplot(data, aes(x = Gender, fill = factor(Recurred))) +
  geom_bar(position = "fill") +
  labs(y = "Proportion", fill = "Recurrence")
```



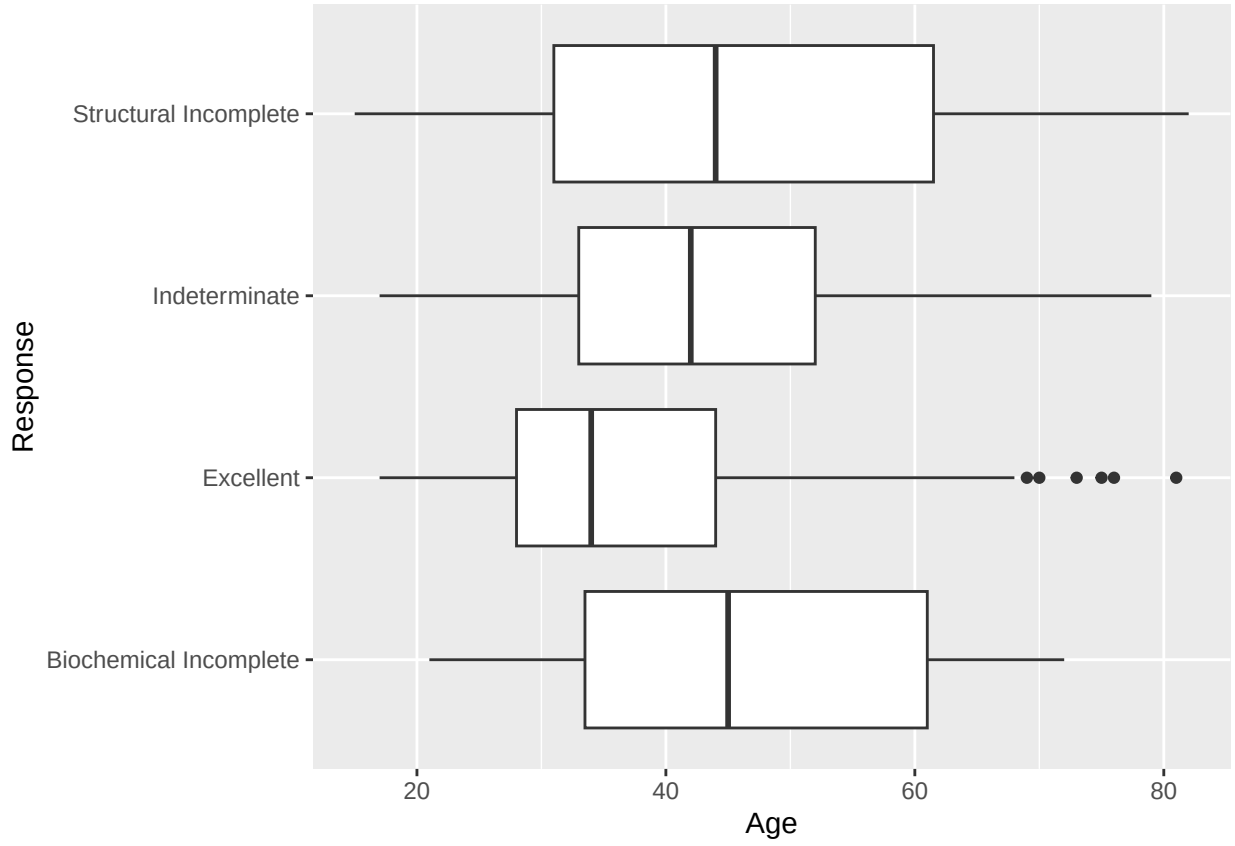

```
ggplot(data, aes(x = factor(Gender), fill = risk_binary)) +
  geom_bar(position = "fill")
```




```
ggplot(data, aes(x = Hx.Radiothreapy, y = Age)) +
  geom_boxplot()
```




```
ggplot(data, aes(x = Age, y = Response)) +
  geom_boxplot()
```




```
model <- aov(Age ~ Response, data)
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Age ~ Response, data = data)
##
## $Response
##                                                 diff          lwr       upr
## Excellent-Biochemical Incomplete             -9.585911 -17.88268942 -1.289133
## Indeterminate-Biochemical Incomplete         -4.066287 -13.30496547  5.172392
## Structural Incomplete-Biochemical Incomplete -0.627807  -9.43965258  8.184039
## Indeterminate-Excellent                       5.519625   0.02195541 11.017294
## Structural Incomplete-Excellent               8.958104   4.21260089 13.703608
## Structural Incomplete-Indeterminate           3.438480  -2.80943739  9.686397
##                                                  p adj
## Excellent-Biochemical Incomplete             0.0160853
## Indeterminate-Biochemical Incomplete         0.6676276
## Structural Incomplete-Biochemical Incomplete 0.9977920
## Indeterminate-Excellent                      0.0486665
## Structural Incomplete-Excellent              0.0000097
## Structural Incomplete-Indeterminate          0.4875162
```

0.4 Discussion

0.5 References