Modeling Thyroid Cancer Recurrence: A Multiple Logistic Regression Approach

Abstract

Thyroid cancer is generally considered to be one of the more treatable forms of cancer. However, when not treated with proper care, certain types and stages can become aggressive, difficult to treat, and potentially life-threatening. This study investigates the predictive power of initial treatment outcome on thyroid cancer recurrence using a multiple logistic regression model, as well as univariate and exploratory analyses of additional variables. Our findings confirm that treatment outcome is a strong and statistically significant predictor of recurrence, after controlling for age, gender, prior radiotherapy, and risk classification. While age and gender were not statistically significant in the multivariable model, univariate and exploratory analyses suggest individual associations with recurrence and influence on treatment outcome. However, the observational nature of the study and lack of random sampling limits its capacity for causal inference and generalizability.

1 Introduction

Cancer, characterized by the uncontrolled proliferation of cells, remains one of the leading cause of death around the world [1, 2]. While many cancer types have observed declining mortality rates, thyroid cancer has seen a slight upward trend and 44,000 new cases are projected for this year [3]. Although often treatable through surgery or Radioactive Iodine (RAI) therapy, treatment outcome varies [4]. This makes understanding the relationship between initial treatment outcome and the likelihood of recurrence essential for effective clinical follow-up and subsequent interventions [5].

This study investigates the following research question: Controlling for age, gender, prior radiotherapy, and clinical risk classification, is initial treatment outcome a statistically significant predictor of thyroid cancer recurrence? We hypothesized that the initial treatment outcome is a statistically significant predictor of thyroid cancer recurrence, holding other variables we suspect to be significant constant. Our findings indicate that treatment outcome is a statistically significant predictor of recurrence when controlling for age, gender, radiotherapy history, and risk classification. We also explored the influence and interrelationships of other explanatory variables to provide a more comprehensive understanding of recurrence risk for thyroid cancer patients.

2 Method

The data for this study was obtained from a publicly available dataset on Kaggle [6]. This dataset originates from a published article by Hamadan University in the European Archives of Oto-Rhino Laryngology [7]. It comprises retrospective data from 383 thyroid cancer patients treated at the Hamadan University of Medical Sciences in Iran. Each patient was followed for a minimum of 10 years, from initial diagnosis all the way through the time of surgery, with records spanning a 15-year period. The dataset includes information related to patient demographics, treatment history, and clinical outcomes.

For detailed analysis, we recoded the Recurrence variable as a numeric binary indicator (i.e., 0 = No Recurrence, 1 = Recurrence) and collapsed Risk into a binary variable (i.e., High vs. Not High). We then fit a multiple logistic regression model using recurrence as the response variable and treatment outcome, age, gender, radiotherapy history, and binary risk classification as the predictors. All other variables in the dataset were eliminated for simplicity.

Since our primary predictors are categorical, the assumption of linearity in the logit is satisfied by design. For the sole quantitative variable, Age, we constructed an empirical logit plot to visually confirm that the assumption of linearity has been met. In Figure 1 in the Appendix, we can see that variance remains relatively constant with the exception of the 4th quintile. However, it is not outside the standard deviation, and the conditions for linearity seem to be relatively satisfied. The assumption of independence was reasonably satisfied, as each patient was observed and followed individually with no repeated measures or clustering. The assumption of randomness, however, was only partially met. Although patients were randomly assigned to training and validation sets, they were originally drawn from a single medical center, limiting the generalizability of the results. As such, inferential statistics should be interpreted as valid within this clinical sample, but not necessarily generalizable to the broader thyroid cancer population without further external validation.

3 Results

3.1 Multiple Logistic Regression Model

We fit the logistic regression model described above to assess the relationship between initial treatment outcome and thyroid cancer recurrence. The full model coefficients, standard errors, z-values, and p-values

for each term are summarized in Table 1 in the Appendix. The fitted model is structured as follows, with "Indeterminate" as the reference category for the TreatmentOutcome variable:

$$\begin{aligned}\log\left(\frac{\pi}{1-\pi}\right) = 22.392 \\ &- 3.034\,(\text{TreatmentOutcomeExcellent}) \\ &+ 1.663\,(\text{TreatmentOutcomeBiochemicalIncomplete}) \\ &+ 5.671\,(\text{TreatmentOutcomeStructuralIncomplete}) \\ &+ 0.030\,(\text{Age}) \\ &+ 1.043\,(\text{GenderM}) \\ &- 10.634\,(\text{RadiotherapyYes}) \\ &- 26.058\,(\text{risk\_binaryNotHigh})\end{aligned}$$

As hypothesized, TreatmentOutcome was a strong and statistically significant predictor of thyroid cancer recurrence. Specifically, "Excellent" had a p-value = 0.005, "Biochemical Incomplete" had a p-value = 0.009, and "Structural Incomplete" had a p-value < 0.001. The direction and magnitude of these effects are consistent with clinical expectations: a more favorable initial treatment outcome is associated with a lower risk of recurrence. See Figure 2 for the visualization.

While Age (p-value = 0.125) and Gender (p-value = 0.107) were not statistically significant predictors at the conventional $\alpha = 0.05$ level in the multivariable model, their p-values are close to the significance level of $\alpha = 0.1$ and may be worth investigating. Notably, holding all else constant, the odds of recurrence increases by approximately 3% for each additional year of age, as $e^{0.03} = 1.03$.

3.2 Univariate and Exploratory Analyses

3.2.1 Univariate Logistic Regression

To isolate the individual effects of Age and Gender, we performed separate univariate logistic regressions. (See Tables 2 and 3). Interestingly, when modeled individually, both Age and Gender were statistically significant predictors of thyroid cancer recurrence. Age had a p-value = 9.82e-07, and Gender had a p-value = 1.39e-09. However, as we saw earlier, neither Age nor Gender were statistically significant at the 0.05 level in the multivariable model. This discrepancy between their significance in the univariate and multivariable models suggests potential confounding variables or multicollinearity. That is, the effects of Age and Gender may be explained or diminished once other covariates, such as treatment outcome or clinical risk classification, are taken into account.

3.2.2 Exploratory Plots

Further exploratory analyses provided visual insights into the relationships between Age, Gender, and Recurrence. Figure 3 shows an upward trend in the logistic regression curve for age, indicating a positive association between age and thyroid cancer recurrence. Figure 4 displays a proportional bar chart comparing recurrence rates by gender. Visually, male patients exhibit a higher proportion of recurrence compared to female patients, implying that gender may also play a role in recurrence risk.

We also explored whether Age and Gender are associated with Treatment Outcome, which could help explain the observed discrepancies. Figure 5 reveals that patients in the "Excellent" treatment outcome group tend to be younger, with a lower median age compared to the other groups. The remaining categories exhibit more similar age distributions with greater spread and higher medians, suggesting a potential inverse relationship between younger age and more favorable treatment outcomes.

Similarly, Figure 6 illustrates that Gender also appears to influence treatment response. Female patients are disproportionately more likely to achieve an "Excellent" treatment response, while male patients appear

disproportionately more likely to fall into the "Structural Incomplete" group. The "Indeterminate" and "Biochemical Incomplete" categories are more evenly distributed across genders.

3.2.3 One-Way ANOVA

To formally assess the trends in Age across treatment outcome groups, we conducted a One-Way ANOVA. (See Table 4). The conditions for this model were sufficiently satisfied, supporting the validity of our analysis. (See Figure 7). The ANOVA revealed statistically significant differences in Age across treatment outcome groups. We followed with a Tukey's post-hoc test to analyze the pairwise differences. (See Table 5). These tests revealed that the "Excellent" group were statistically significantly different from each of the other categories. This finding reinforces the visual trend observed in Figure 5 and confirms the inverse relationship between age and favorable treatment outcome.

3.2.4 Multicollinearity Assessment

Although these analyses are exploratory and descriptive in nature, it provides initial evidence that both Age and Gender may influence treatment outcomes, which in turn could affect recurrence risk. Given these relationships, we considered the potential for multicollinearity, particularly among Age, Gender, and Treatment Outcome. (See Table 6). Interestingly, the adjusted GVIF values were well below the conventional thresholds of 5 to 10, indicating that multicollinearity is not a serious concern in this model. We also examined pairwise Pearson correlations among numeric predictors. (See Table 7). Radiotherapy and Risk exhibit a moderate correlation (0.38), which aligns with the slightly elevated GVIFs for both variables. This relationship is clinically plausible, as high-risk patients are more likely to receive radiotherapy. All other correlations were weak to moderate and did not indicate concern for problematic multicollinearity.

4 Discussion

Our results support the hypothesis that treatment outcome is a strong and statistically significant predictor of thyroid cancer recurrence. Specifically, patients classified with a "Structural Incomplete" response to initial treatment exhibited significantly higher proportions of recurrence. When examining other predictors, Age and Gender were not statistically significant in the multiple logistic regression model. However, univariate and exploratory analyses indicated that they are individually associated with recurrence and may influence treatment outcome. Assessment of multicollinearity revealed weak to moderate levels among the explanatory variables, but not at levels that would invalidate our model.

However, our findings must be interpreted within the context of the study's inherent limitations. For one, the dataset was not derived from a randomly sampled population of thyroid cancer patients. Instead, it was all sourced from the Hamadan University of Medical Sciences in Iran. While each patient was independently observed within Hamadan University, the fact that the sample was all sourced from a single medical center restricts the generalizability of our conclusions to the broader thyroid cancer patient population. Secondly, given the observational design of this study, causal inference cannot be made. We can conclude that treatment outcome is strongly associated with recurrence, but we cannot conclude that one causes the other. Furthermore, both Risk and Radiotherapy had large standard errors due to uneven distributions within their respective categories. Though not statistically significant predictors in this model, their relationship remains clinically important and warrants cleaner data or larger sample sizes to clarify their roles.

Future research should consider including additional clinical variables such as tumor size, molecular markers, or comorbidities that may help refine the model. Further studies should also consider experimental designs to enable causal inference, and use a more diverse sample for generalizability.

5 References

[1] Brown, J. S., Amend, S. R., Austin, R. H., Gatenby, R. A., Hammarlund, E. U., & Pienta, K. J. (2023). Updating the Definition of Cancer. Molecular cancer research : MCR, 21(11), 1142–1147. https://doi.org/10.1158/1541-7786.MCR-23-0411

[2] Nagai, H., & Kim, Y. H. (2017). Cancer prevention from the perspective of global cancer burden patterns. Journal of thoracic disease, 9(3), 448–451. https://doi.org/10.21037/jtd.2017.02.75

[3] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 74(3), 229–263. https://doi.org/10.3322/caac.21834

[4] Sun, Y. D., Zhang, H., Zhu, H. T., Wu, C. X., Chen, M. L., & Han, J. J. (2022). A systematic review and meta-analysis comparing tumor progression and complications between radiofrequency ablation and thyroidectomy for papillary thyroid carcinoma. Frontiers in oncology, 12, 994728. https://doi.org/10.3389/fonc.2022.994728

[5] Kim, J. Y., Myung, J. K., Kim, S., Tae, K., Choi, Y. Y., & Lee, S. J. (2024). Prognosis of Poorly Differentiated Thyroid Carcinoma: A Systematic Review and Meta-Analysis. Endocrinology and metabolism (Seoul, Korea), 39(4), 590–602. https://doi.org/10.3803/EnM.2024.1927

[6] Thyroid Cancer Recurrence Dataset. Kaggle. (2025). https://www.kaggle.com/datasets/aneevinay/thyroid-cancer-recurrence-dataset

[7] Borzooei, S., Briganti, G., Golparian, M. et al. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. Eur Arch Otorhinolaryngol 281, 2095–2104 (2024). https://doi.org/10.1007/s00405-023-08299-w

6 Appendix

6.1 Tables

Table 1: Variable Descriptions

Variable	Description	Levels
Recurrence	Whether thyroid cancer recurred.	Yes or No
Age	Patient's age, recorded in years.	N/A
Gender	Biological sex.	Male or Female
Radiotherapy	History of prior radiotherapy.	Yes or No
Risk	Cancer risk classification.	Low, Medium, High
TreatmentOutcome	Initial treatment outcome.	Excellent, Indeterminate, Structural Incomplete, Biochemical Incomplete

Table 2: Multiple Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	22.392	3357.120	0.007	0.995
TreatmentOutcomeExcellent	-3.034	1.089	-2.786	0.005
TreatmentOutcomeBiochemical Incomplete	1.663	0.637	2.612	0.009
TreatmentOutcomeStructural Incomplete	5.671	0.875	6.478	0.000
Age	0.030	0.020	1.532	0.125
GenderM	1.043	0.646	1.614	0.107
RadiotherapyYes	-10.634	2414.759	-0.004	0.996
risk_binaryNot High	-26.058	3357.119	-0.008	0.994

Table 3: Univariate Logistic Regression: Recurred ~ Age

term	estimate	std.error	statistic	p.value
(Intercept)	-2.5227446	0.3550921	-7.10448	1.21e-12
Age	0.0373481	0.0076294	4.89525	9.82e-07

Table 4: Univariate Logistic Regression: Recurred ~ Gender

term	estimate	std.error	statistic	p.value
(Intercept)	-1.315677	0.1386237	-9.490993	< 2e-16
GenderM	1.686051	0.2784041	6.056128	1.39e-09

Table 5: One-Way ANOVA: Age by Treatment Outcome

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TreatmentOutcome	3	6357.015	2119.0050	9.897598	2.7e-06
Residuals	379	81141.194	214.0929	NA	NA

Table 6: Tukey HSD: Age by Treatment Outcome

Comparison	Difference	Lower CI	Upper CI	Adjusted p-value
Excellent-Indeterminate	-5.520	-11.017	-0.022	0.049
Biochemical Incomplete-Indeterminate	4.066	-5.172	13.305	0.668
Structural Incomplete-Indeterminate	3.438	-2.809	9.686	0.488
Biochemical Incomplete-Excellent	9.586	1.289	17.883	0.016
Structural Incomplete-Excellent	8.958	4.213	13.704	0.000
Structural Incomplete-Biochemical Incomplete	-0.628	-9.440	8.184	0.998

Table 7: Generalized Variance Inflation Factors (GVIFs)

	GVIF	Df	GVIF^(1/(2*Df))
TreatmentOutcome	1.120	3	1.019
Age	1.105	1	1.051
Gender	1.032	1	1.016
Radiotherapy	2.068	1	1.438
risk_binary	2.068	1	1.438

Table 8: Correlation Matrix of Numeric Predictors

	Age	Gender	Radiotherapy	Risk
Age	1.00	0.19	0.18	0.29
Gender	0.19	1.00	0.24	0.22
Radiotherapy	0.18	0.24	1.00	0.38
Risk	0.29	0.22	0.38	1.00

6.2 Figures

Figure 1: Empirical Logit Plot

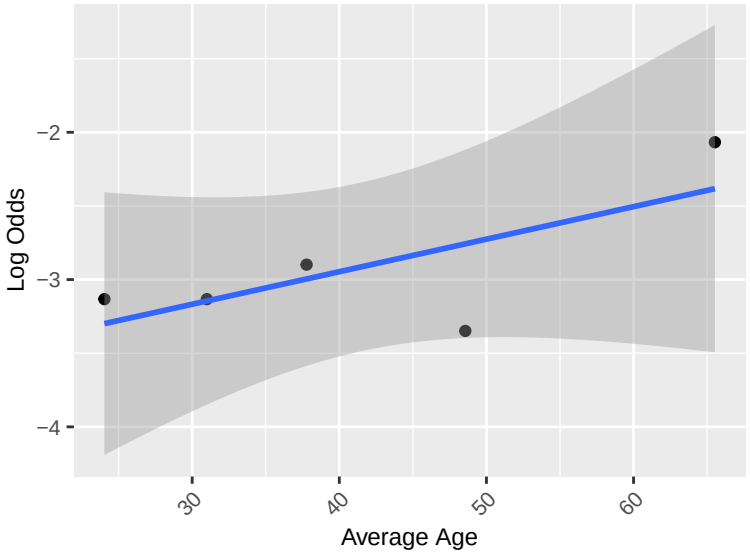

Figure 2: Proportion of Recurrence by Treatment Response

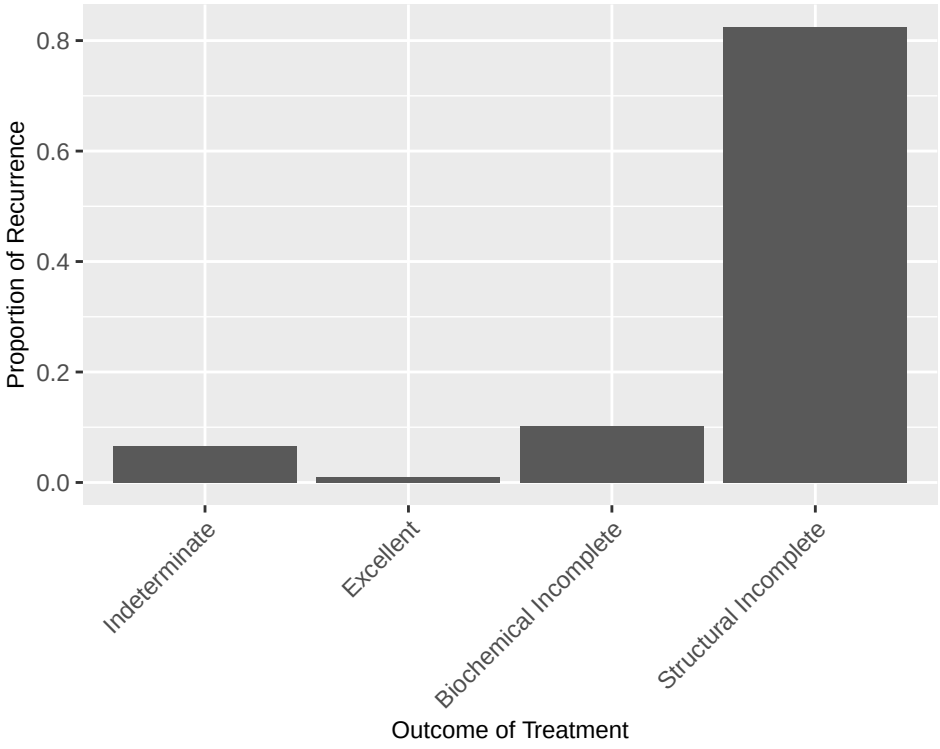

Figure 3: Logistic Regression of Recurrence by Age

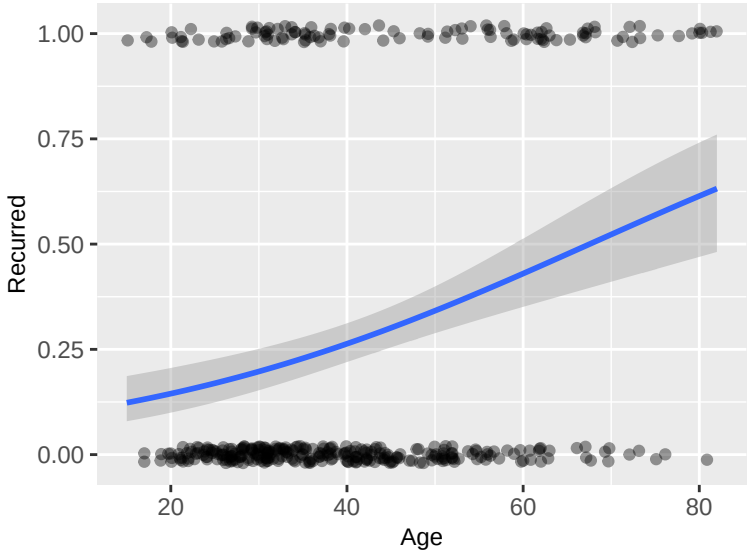

Figure 4: Stacked Bar Chart of Recurrence by Gender

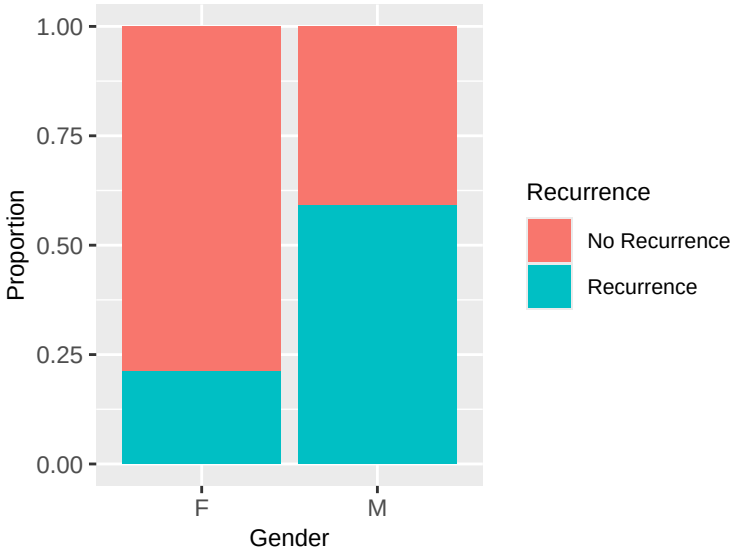

Figure 5: Boxplot of Treatment Outcome by Age

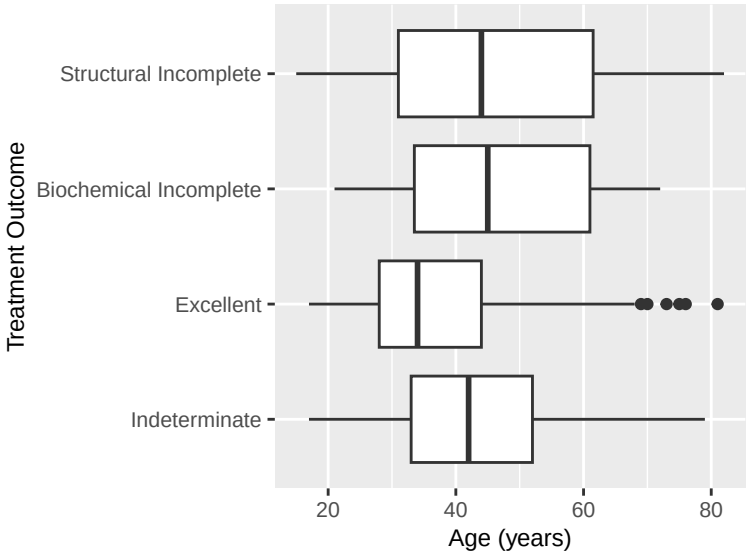

Figure 6: Stacked Bar Chart of Treatment Outcome by Gender

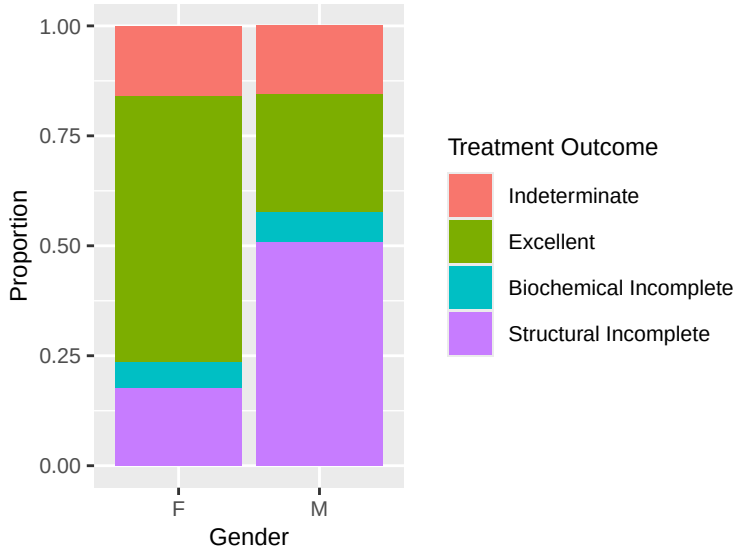

Figure 7: Diagnostic Plots of Age Across Treatment Outcome Groups