Modeling of Serum Total Cholesterol Levels from NHANES 2017 to 2020

Qianxuan Huang, qh2285

Abstract

Cholesterol is a crucial risk factor for cardiovascular diseases which in their turn are among the main causes of death worldwide and public health concern, with heart diseases being the most prevalent ones. For cholesterol control, the early prediction is considered one of the most effective ways. This study explores the prediction of serum total cholesterol level using two modeling approaches: linear regression model and generalized additive model, with gender as the primary predictor and age, race, marital status, education level, poverty level, BMI, physical activity level, sleep duration, alcohol use level, alcohol drinks order, and cotinine level as covariates.

Findings suggest that the GAM slightly outperforms the linear model in capturing cholesterol variability. However, neither model provides strong predictive power in this dataset. These results highlight the potential of machine learning techniques for cholesterol prediction, while also underscoring the need for early intervention in large-scale public health efforts. Future research should investigate additional factors and interaction effects to improve model accuracy and clinical utility.

Introduction

In this study, we want to predict serum total cholesterol levels in adults by predictor gender and other potential covariance. Supported by the previous research, the main risk factors potentially impacting on the cholesterol levels include gender, age, BMI, physical activity level, alcohol

level (Fazakis, 2021), poverty level, education level (Lara, 2018), sleep hour (Confortin, 2022), race, married status (Smiley, 2019), and cotinine (Emberson, 2003).

In previous studies, a linear regression modeling planned to assess the associations between total cholesterol, and predictors such as age, sex, race, BMI, physical activity, alcohol consumption, education level, and serum cotinine levels. The study provided a framework for understanding how these variables interact to influence cholesterol levels in adults by data from NHANES 2003–2006 (Sternberg, 2013). A machine learning technique was used to identify individuals at risk and facilitate earlier intervention to prevent the future development of cholesterol, by unutilized the English Longitudinal Study of Ageing (ELSA), a dataset was derived to evaluate the long-term cholesterol risk of elderly people. Predictor in the study included gender, age, BMI, physical activity level, and alcohol level (Fazakis, 2021). Another machine learning technique of Generalized additive models (GAM) has been used to identify the smooth relationship between the HDL cholesterol/ triglyceride/ LDL cholesterol, and the sleep duration, by included 2705 participants from The National Health and Nutrition Examination Survey, 2013/2014. Models were adjusted for age, sex, race, marital status, household size, sitting time and physical activity. The result was that there was no significant non-linear association between sleep duration and LDL cholesterol in GAM (Smiley, 2019).

Methods

The resource is from the program The National Health and Nutrition Examination Survey (NHANES) conducted by the CDC from 2017 to March 2020 Pre-pandemic. It combines interviews and physical examinations to collect a wide range of health-related information. This

program divides different types of measurements into sub-dataset. And the dataset used in this study is a combination of 8 raw xpt files of examination, laboratory, and questionnaire datasets.

The outcome of this study is serum total cholesterol levels (mg/dL) to assess the risk of high cholesterol with gender as predictor. Age, race, marital status, education level, poverty level, BMI, physical activity level, daily sleep hour, alcohol use level, alcohol drinks order (1-11 from least to most drinking pattern), and serum cotinine level (ng/mL) as potential covariates. After deleted all participants who had missing data and younger than 18 years old, the cleaned dataset involves 5374 with 11 characteristics and cholesterol related variables (Table 1). The analysis is conducted using RStudio (version 4.4.1).

In the first method, a linear regression model is built to predict serum total cholesterol levels. Dummy variables are created for categorical variables. The reference group for gender, race, marital status, education level, poverty level, physical activity level, alcohol use level are male, other race, never married, less than 9 years, below 130% of poverty guidelines, light/unknow activity, light drinker respectively. R package tidyverse is used for data manipulation (dplyr), visualization (ggplot2), and data import (readr). R package gtsummary and flextable are used to generate and format model summaries and regression tables for reporting. R package broom is used for converting statistical analysis objects into tidy tibbles.

In the second method, a non-linear model is built to predict serum total cholesterol levels. A machine learning method of Generalized Additive Models (GAM) allows for flexible relationships, which may capture complex patterns in the data. Categorical variables are composed of discrete levels, and therefore applying smooth terms to them is not meaningful, as they do not represent continuous quantities that can vary smoothly. In GAM, such variables are automatically treated as dummy variables by R and included in the linear portion of the model

rather than smooth functions. The continuous variables exhibiting non-linear relationships with the outcome are modeled using smooth terms. The dataset splits into 80% training and 20% testing data automatedly by R. R package tidymodels is used for structuring the modeling workflow. R package caret is used to train the GAM with 10-fold cross-validation and for tuning hyperparameters (e.g., selection method and smoothing parameters). R package mgcv is used to underpin the GAM fitting process, enabling smooth term estimation via penalized regression splines. R package broom and ggplot2 are also used for interpreting and visualizing the results.

Results

The overall of data is summary in Table 1. Gender as the predictor in the model is approximately evenly distributed, with 49.6% female and 50.4% male participants. This balanced distribution suggests that the model is not biased toward either gender.

Table 1: Baseline Characteristics

Characteristic	**N = 5,374**[1]
serum total cholesterol levels (mg/dL)	
Mean (SD)	186 (42)
Median (Q1 - Q3)	182 (157 - 211)
Min - Max	71 - 446
age	
Mean (SD)	51 (17)
Median (Q1 - Q3)	52 (36 - 64)
Min - Max	20 - 80

Characteristic	**N = 5,374**[1]
gender	
Female	2,667 (50%)
Male	2,707 (50%)
race	
Hispanic	1,118 (21%)
Non-Hispanic Asian	480 (8.9%)
Non-Hispanic Black	1,327 (25%)
Non-Hispanic White	2,178 (41%)
Other Race	271 (5.0%)
marital status	
Married/Living with Partner	3,164 (59%)
Never married	986 (18%)
Widowed/Divorced/Separated	1,224 (23%)
education level	
9-11th grade	550 (10%)
College graduate or above	1,383 (26%)
High school graduate/GED or equivalent	1,286 (24%)
Less than 9th grade	273 (5.1%)
Some college or AA degree	1,882 (35%)
poverty level	
Above 185% of Poverty Guidelines	2,914 (54%)
Below 130% of Poverty Guidelines	1,629 (30%)
Between 130% and 185% of Poverty Guidelines	831 (15%)
BMI	

Characteristic	**N = 5,374**[1]
Mean (SD)	30 (8)
Median (Q1 - Q3)	29 (25 - 34)
Min - Max	15 - 92
physical activity level	
Light/Unknown activity	1,271 (24%)
Moderate activity	1,761 (33%)
Vigorous activity	2,342 (44%)
daily sleep hour	
Mean (SD)	7.52 (1.62)
Median (Q1 - Q3)	7.50 (6.50 - 8.50)
Min - Max	2.00 - 14.00
alcohol use level	
Heavy Drinker	775 (14%)
Light Drinker	2,590 (48%)
Moderate Drinker	2,009 (37%)
alcohol use order	
Mean (SD)	5 (3)
Median (Q1 - Q3)	5 (2 - 7)
Min - Max	1 - 11
serum cotinine level (ng/mL)	
Mean (SD)	61 (134)
Median (Q1 - Q3)	0 (0 - 16)
Min - Max	0 - 1,620

[1]n (%)

Method I:

In the first method, all predictor and potential covaries are used to build a general a linear regression model to predict serum total cholesterol levels (Table 2). Using the original total cholesterol levels as the outcome resulted in very small coefficient estimates in the model. To address this issue and improve interpretability, a log transformation is applied and used log(total cholesterol levels) as the outcome variable.

Table 2: General Linear Model

term	estimate	std.error	p.value	conf.low	conf.high
Intercept	5.1542	0.0315	<0.0001	5.0924	5.2160
Age	0.0003	0.0002	0.1832	-0.0001	0.0007
Gender: Male (ref)					
Female	0.0574	0.0063	<0.0001	0.0450	0.0698
Race: Other Race (ref)					
Hispanic	-0.0113	0.0151	0.4566	-0.0409	0.0184
Non-Hispanic White	-0.0237	0.0142	0.0946	-0.0516	0.0041
Non-Hispanic Black	-0.0477	0.0146	0.0011	-0.0764	-0.0190
Non-Hispanic Asian	0.0095	0.0171	0.5798	-0.0241	0.0430
Marital Status: Never Married (ref)					
Married/Living with Partner	0.0400	0.0086	<0.0001	0.0231	0.0569
Widowed/Divorced/Separated	0.0539	0.0105	<0.0001	0.0333	0.0745
Education Level: Less than 9th grade (ref)					
Education 9-11th grade	-0.0300	0.0166	0.0712	-0.0626	0.0026
High school graduate/GED or equivalent	-0.0403	0.0153	0.0086	-0.0704	-0.0103
Some college or AA degree	-0.0255	0.0152	0.0941	-0.0553	0.0044

term	estimate	std.error	p.value	conf.low	conf.high
College graduate or above	-0.0132	0.0160	0.4099	-0.0445	0.0181
Poverty Level: Below 130% of Poverty Guidelines (ref)					
Between 130% and 185% of Poverty Guidelines	0.0164	0.0094	0.0822	-0.0021	0.0349
Above 185% of Poverty Guidelines	0.0127	0.0076	0.0969	-0.0023	0.0276
Physical Activity Level: Light/Unknown activity (ref)					
Vigorous physical activity	0.0039	0.0082	0.6382	-0.0122	0.0199
Moderate physical activity	0.0056	0.0081	0.4882	-0.0103	0.0216
Alcohol Used Level: Light Drinker (ref)					
Moderate alcohol drinker	-0.0201	0.0143	0.1595	-0.0481	0.0079
Heavy alcohol drinker	-0.0060	0.0241	0.8037	-0.0533	0.0413
BMI	-0.0001	0.0004	0.7628	-0.0009	0.0007
Sleep hour	-0.0039	0.0019	0.03545	-0.0076	-0.0003
Alcohol use order	0.0094	0.0029	0.0012	0.0037	0.0151
serum cotinine level	-0.00001	0.00002	0.5819	-0.0001	0.00004

Based on statistical significance at the $\alpha = 0.05$ level, non-significant covariates were removed, resulting in only five covaries and predictor, which are gender, marital status, education level, sleep hours, and alcohol use order. A new reduced linear model was then constructed using these variables. However, in this reduced model, the p-values for sleep hours and college graduate or above are above 0.05, suggesting a potential lack of significance. To further evaluate the contribution of education level, a partial F-test was conducted in Table 3, comparing models with and without the education level variables. The result indicated that education level remained a

significant predictor ($p < 0.0001$). Therefore, the final linear model retained education level and excluded sleep hours, as presented in Table 4.

Table 3: Partial F-test of Reduced Linear Model without Education Level

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
5,369	259.1776				
5,364	257.7933	5	1.384302	5.760738	<0.0001

Table 4: Final Linear Model

term	estimate	std.error	p.value	conf.low	conf.high
Intercept	5.1232	0.0158	<0.0001	5.0922	5.1542
Gender: Male (ref)					
Female	0.0545	0.0061	<0.0001	0.0425	0.0666
Marital Status: Never Married (ref)					
Married/Living with Partner	0.0528	0.0080	<0.0001	0.0370	0.0685
Widowed/Divorced/Separated	0.0625	0.0094	<0.0001	0.0440	0.0809
Education Level: Less than 9th grade (ref)					
Education 9-11th grade	-0.0395	0.0163	0.0152	-0.0714	-0.0076
High school graduate/GED or equivalent	-0.0498	0.0147	0.0007	-0.0786	-0.0210
Some college or AA degree	-0.0322	0.0143	0.0247	-0.0604	-0.0041
College graduate or above	-0.0121	0.0147	0.4105	-0.0409	0.0167
Alcohol use order	0.0074	0.0010	<0.0001	0.0054	0.0093

To visualize the final linear model, a Q-Q plot shows moderate departures from normality in the residuals in Figure1. The departure from normality in the residual plot is very minor and not a big concern. Figure 2 demonstrates a systematic alignment between predicted and observed values. While the model captures central trends, the narrow range of predictions relative to the

broader spread of true values suggests potential underfitting or limited explanatory power. A perfect model would show points aligned diagonal, which is the black line in Figure 2.

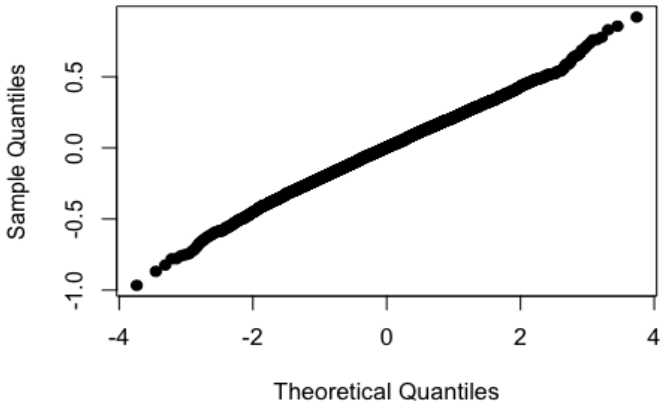

Figure 1: Q-Q Plot of Residuals in Linear Model

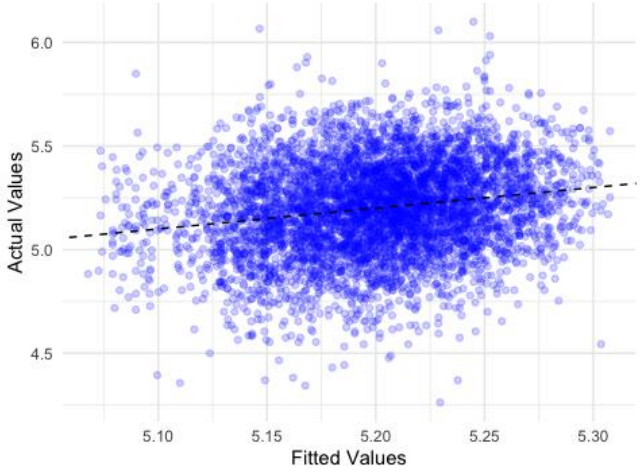

Figure 2: Fitted comparison of Actual Value in Linear Model

Method II:

In the second method, all predictor and potential covaries are used to build a GAM to predict log transformation of serum total cholesterol levels. Table 5 shows the linear term in the general model and Table 6 shows the non-linear smooth terms with coefficients and p-value. The general formula is log(total serum cholesterol level) ~ gender + marital status + poverty level + physical

activity level + alcohol use level + race + education level + s(alcohol use order) + s(sleep hour) + s(age) + s(BMI) + s(serum cotinine level).

Table 5: Linear Term of General GAM

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	5.2567	0.0220	239.2899	<0.0001
Gender: Female (ref)				
Male	-0.0575	0.0069	-8.3591	<0.0001
Marital status: Married/Living with Partner (ref)				
Never married	-0.0067	0.0095	-0.7039	0.4815
Widowed/Divorced/Separated	0.0246	0.0084	2.9065	0.0037
Poverty level: Above 185% of Poverty Guidelines (ref)				
Below 130% of Poverty Guidelines	-0.0181	0.0083	-2.1934	0.0283
Between 130% and 185% of Poverty Guidelines	0.0112	0.0097	1.1579	0.2470
Physical activity: Light/Unknown activity (ref)				
Moderate activity	-0.0034	0.0088	-0.3854	0.6999
Vigorous activity	-0.0052	0.0088	-0.5831	0.5598
Alcohol use level: Heavy alcohol drinker (ref)				
Light Drinker	-0.0057	0.0245	-0.2341	0.8149
Moderate Drinker	-0.0168	0.0141	-1.1907	0.2338
Race: Hispanic (ref)				
Non-Hispanic Asian	0.0280	0.0140	2.0050	0.0450
Non-Hispanic Black	-0.0370	0.0102	-3.6217	0.0003
Non-Hispanic White	0.0036	0.0095	0.3754	0.7074
Other Race	0.0179	0.0161	1.1085	0.2677
Education level: Education 9-11th grade (ref)				
College graduate or above	-0.0095	0.0130	-0.7262	0.4678
High school graduate/GED or equivalent	-0.0216	0.0122	-1.7691	0.0770
Less than 9th grade	0.0034	0.0183	0.1852	0.8531
Some college or AA degree	-0.0113	0.0119	-0.9478	0.3433

Table 5: Non-linear Smooth Term of General GAM

	edf	Ref.df	F	p-value
s(alcohol use order)	0.8633	8	0.7880	0.0069
s(sleep hour)	0.0002	9	0.0000	0.7446
s(age)	4.9732	9	26.5599	<0.0001
s(BMI)	3.5974	9	3.9277	<0.0001
s(serum cotinine level)	0.8426	9	0.5900	0.012

However, none of the covariates included as linear terms were statistically significant at the $\alpha = 0.05$ level, except for the predictor gender. In contrast, the smooth terms for alcohol use order, age, BMI, and cotinine were found to be significant. Based on these findings, a reduced GAM was constructed incorporating only the significant predictors. Table 7 shows the linear term in the general model and Table 8 shows the non-linear smooth terms with coefficients and p-value. The visualization plots in Figure 3 revealed significant non-linear effects. The reduced formula is log(total serum cholesterol level) ~ gender + s(alcohol use order) + s(age) + s(BMI) + s(serum cotinine level). The testing data are used to calculate adjusted R^2 and RMSE to the comparison of models in discussion.

Table 7: Linear Term in Reduced GAM

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	5.2298	0.0047	1120.3471	<0.0001
Gender: Female (ref)				
Male	-0.0586	0.0067	-8.7953	<0.0001

Table 8: Non-Linear Smooth Term in Reduced GAM

	edf	Ref.df	F	p-value
s(alcohol use order)	1.0037	1.0075	39.1517	<0.0001
s(age)	5.5275	6.6548	41.0858	<0.0001
s(BMI)	4.6315	5.6971	8.0713	<0.0001
s(serum cotinine level)	1.0303	1.0599	9.2637	0.0019

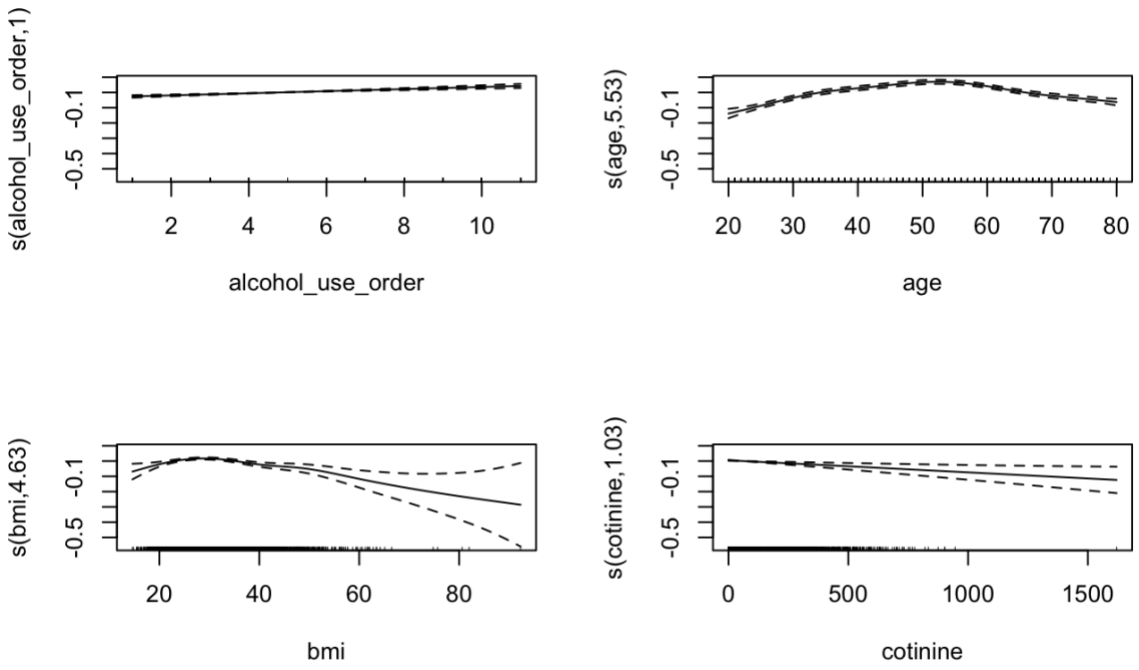

Figure 3: Estimated Smooth Terms with 95% CI in Reduced GAM

Discussion

Based on the estimated coefficient of predictor gender, females' total serum cholesterol levels are higher than those of males in both models. Females have a 0.0545-unit (5.6%) higher log-transformed total serum cholesterol level in linear model, while males have a 0.0586-unit (5.7%) lower log-transformed level. Such conclusion is supported by published literature that women (185.5 mg/dL) had higher average total cholesterol and LDL cholesterol levels than men (169.2 mg/dL), which is approximately 8.8% (Gupta, 2016). In addition, alcohol used order is the only covaries appeared in both models that higher order leading to higher total serum cholesterol levels.

RMSE of final linear model is 0.2191 and R^2 is 0.0359, meaning that this model is not a good fit of the dataset. The GAM yields an RMSE of 0.2176 and an R^2 of 0.0698, indicating that it

explains only a limited proportion of the variance in total cholesterol and may not provide a strong overall fit to the dataset. Although GAM achieves a slightly lower RMSE value, both RMSE values are very similar. In addition, the GAM achieves a higher R^2, suggesting it captures more of the variance in total cholesterol. We selected the GAM for further interpretation, while retaining the linear model as a baseline.

One major advantage of GAM approach is its ability to capture non-linear trends in continuous variables using smooth functions, which may lead to more accurate predictions. Moreover, GAM can automatically decide the degree of smoothness, thus providing a data-driven way to balance bias and variance. However, GAM is not a good method for categorical variables. The natural interpretability of GAM breaks down when there are more than two classes (Zhang, X., 2019), with is the same situation in our analysis. Additionally, smooth functions may result in overfitting, especially covaries BMI has a wide range of 95% CI in large value.

In contrast, the reduced linear model is considerably more straightforward and interpretable. This simplicity facilitates clear communication of the effects to stakeholders and offers ease of computation and inference, with well-established methods for hypothesis testing and model diagnostics. On the downside, the reduced linear model might oversimplify the underlying relationships in the data. Additionally, because the model retains fewer covariates than GAM, it might ignore some important non-linear predictors that could enhance the overall model fit.

Future studies should expand the set of candidate predictors and interactions to better capture the variance in total serum cholesterol levels, which may yield improved predictive performance. Thus, research should focus on hybrid modeling strategies that combine the interpretative clarity of linear methods with the flexibility of non-linear smooth functions. Another important consideration is the potential for overfitting observed in some smooth terms. Future research

should examine regularization techniques or alternative smoothing parameter selection methods to mitigate overfitting, particularly in datasets with limited sample sizes or high variability.

References

Fazakis, N., Dritsas, E., Kocsis, O., Fakotakis, N., & Moustakas, K. (2021, October). Long-term Cholesterol Risk Prediction using Machine Learning Techniques in ELSA Database. In IJCCI (pp. 445-450).

Confortin, S. C., Aristizábal, L. Y. G., da Silva Magalhães, E. I., Barbosa, A. R., Ribeiro, C. C. C., Batista, R. F. L., & Silva, A. A. M. D. (2022). Association between sleep duration and cardiometabolic factors in adolescents. BMC Public Health, 22(1), 686.

Lara, M., & Amigo, H. (2018). Association between education and blood lipid levels as income increases over a decade: a cohort study. BMC Public Health, 18, 1-8.

Emberson, J. R., Whincup, P. H., Morris, R. W., & Walker, M. (2003). Re-assessing the contribution of serum total cholesterol, blood pressure and cigarette smoking to the aetiology of coronary heart disease: impact of regression dilution bias. European heart journal, 24(19), 1719-1726.

Sternberg, M. R., Schleicher, R. L., & Pfeiffer, C. M. (2013). Regression modeling plan for 29 biochemical indicators of diet and nutrition measured in NHANES 2003-2006. The Journal of nutrition, 143(6), 948S–56S.

Smiley, A., King, D., Harezlak, J., Dinh, P., & Bidulescu, A. (2019). The association between sleep duration and lipid profiles: the NHANES 2013–2014. Journal of Diabetes & Metabolic Disorders, 18, 315-322.

The National Health and Nutrition Examination Survey (NHANES) 2017-March 2020 Pre-pandemic, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020

Gupta, R., Sharma, M., Goyal, N. K., Bansal, P., Lodha, S., & Sharma, K. K. (2016). Gender differences in 7 years trends in cholesterol lipoproteins and lipids in India: Insights from a hospital database. Indian journal of endocrinology and metabolism, 20(2), 211–218.

Zhang, X., Tan, S., Koch, P., Lou, Y., Chajewska, U., & Caruana, R. (2019, July). Axiomatic interpretability for multiclass additive models. In Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining (pp. 226-234).

OpenAI. (2024). ChatGPT (GPT-4) is used to assist with English grammar correction and language refinement during the writing process.