Impact of Physical, Mental Health, and Lifestyle Factors on Diabetes: Logistic Regression and Hypothesis Testing

Emma Mo[1]
[1]UCLA Fielding School of Public Health

I. INTRODUCTION

Diabetes is a significant public health concern worldwide, affecting millions of individuals and imposing a considerable economic and healthcare burden. The condition is closely linked to various lifestyle factors, such as physical activity, smoking, and physical and mental health status. Understanding the relationship between these factors and the risk of diabetes is essential for early detection, prevention, and intervention strategies. This study utilizes the **Heart Disease and Health Indicators Dataset**, a dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS), which contains health-related survey responses from over 70,000 participants. The dataset includes variables related to physical and mental health, lifestyle habits, and pre-existing conditions, providing an opportunity for comprehensive analysis of diabetes predictors.

Recent studies have highlighted the role of physical and mental health in chronic disease management. Poor physical health days have been shown to correlate with an increased risk of diabetes and cardiovascular diseases, while poor mental health days may exacerbate diabetes management through stress-related behaviors such as overeating or physical inactivity. Similarly, smoking and physical activity have been consistently identified as critical lifestyle factors influencing diabetes risk. However, the interactions among these variables and their combined predictive power for diabetes remain less explored, particularly in large and balanced datasets.

In this study, we aim to investigate how **poor mental health**, **poor physical health**, **smoking**, and **physical activity** influence the risk of diabetes. Using statistical modeling and hypothesis testing, we examine these relationships to determine their significance and predictive power. This research focuses on a descriptive and inferential analysis of the dataset, evaluating the effects of these predictors on diabetes outcomes. By leveraging logistic regression models and performance evaluation metrics, this study provides insights into actionable public health strategies for diabetes prevention and management.

II. METHODS & MATERIALS

A. Data Set

The dataset used in this study is the Heart Disease Health Indicators Dataset, publicly available on Kaggle . It is based on the **Behavioral Risk Factor Surveillance System (BRFSS)**, a retrospective cross-sectional survey conducted by the CDC, including 70,692 participants aged 18 years or older. The dataset provides self-reported information on physical and mental health, lifestyle habits (e.g., smoking, physical activity), and pre-existing conditions such as diabetes.

Key variables selected for this study include poor mental health days and poor physical health days within the last 30 days (continuous variables), smoking status, participation in physical activity (binary variables), and diabetes status (binary outcome). The dataset excludes incomplete responses and participants with missing key health or demographic information. Data management involved cleaning and verifying variable ranges, summarizing continuous variables with descriptive statistics, and binary variables with proportions. Outliers were identified using the interquartile range (IQR) but were retained as they represented valid population variability.

B. Study Population

The study population consists of 70,692 adults aged 18 years or older, with self-reported data on physical and mental health, lifestyle habits, and diabetes status collected through the BRFSS survey. The outcome variable, diabetes status, is binary, with 50% of participants identified as diabetic. Descriptive statistics showed that participants reported an average of 3.75 poor mental health days (SD = 8.16) and 5.81 poor physical health days (SD = 10.06). Approximately 47.5% were smokers, and 70.3% reported engaging in physical activity.

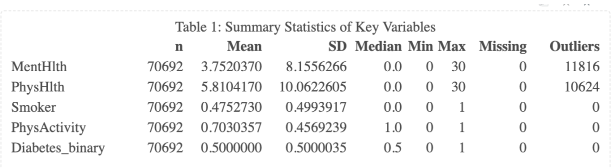

Table 1: Summary Statistics of Key Variables

	n	Mean	SD	Median	Min	Max	Missing	Outliers
MentHlth	70692	3.7520370	8.1556266	0.0	0	30	0	11816
PhysHlth	70692	5.8104170	10.0622605	0.0	0	30	0	10624
Smoker	70692	0.4752730	0.4993917	0.0	0	1	0	0
PhysActivity	70692	0.7030357	0.4569239	1.0	0	1	0	0
Diabetes_binary	70692	0.5000000	0.5000035	0.5	0	1	0	0

C. Statistical Methods

In this study, we used multiple statistical methods to analyze and interpret the relationships between variables and outcomes. For hypothesis testing, Welch's two-sample t-tests were conducted to compare continuous variables, such as poor mental health days (MentHlth) and poor physical health days (PhysHlth), between diabetic and non-diabetic individuals. These tests evaluated whether there were significant differences in the mean values of these variables across the diabetes status groups.

Additionally, Pearson's chi-squared tests were performed to assess the associations between categorical variables, such as smoking status (Smoker) and physical activity (PhysActivity), with diabetes status (Diabetes_binary). These tests evaluated whether the distributions of these lifestyle factors significantly differed across diabetic and non-diabetic groups.

Correlation analysis was used to examine the linear relationship between MentHlth and PhysHlth. A statistically significant positive correlation was observed, indicating that poor mental health days tend to increase with poor physical health days.

The dataset was randomly split into training (50%) and testing (50%) subsets to facilitate model development and evaluation. To model the relationship between diabetes status and its predictors, logistic regression was applied to the training set. The model included MentHlth, PhysHlth, Smoker, and PhysActivity as explanatory variables. Model coefficients and their statistical significance were extracted and interpreted to identify the most influential predictors.

For model evaluation, predictions on the test dataset were used to create a confusion matrix, and the ROC curve was plotted to assess the predictive power of the logistic model. The area under the curve (AUC) was computed as a measure of model performance.

All analyses were conducted using R version 4.4.1. Data manipulation and visualization were facilitated by the dplyr, ggplot2, and pROC libraries. Results were presented in tabular and graphical formats to support interpretation and discussion.

III. RESULTS

A. Hypothesis Testing

Continuous Variables (t-tests):

A two-sample t-test comparing poor mental health days (MentHlth) between diabetic and non-diabetic individuals revealed a statistically significant difference (p-value < 2.2e-16). On average, individuals with diabetes report 4.46 days of poor mental health within the last month, compared with 3.04 days of poor mental health of those without diabetes, which shows that having diabetes is likely to be associated with poor mental health.

A similar t-test for poor physical health days (PhysHlth) also showed a significant difference (p-value < 2.2e-16). Individuals with diabetes report an average of 7.95 days of poor physical health, while those without diabetes have only about 3.67 poor physical health days over the last month. This shows that diabetic individuals experience approximately 4.3 more poor physical health days than non-diabetic individuals.

Categorical Variables (Chi-square Tests):

Among the diabetic sample, 51.8% are smokers, while the proportion of smokers in the non-diabetic group is only 43.2%. The chi-square test for smoking (Smoker) and diabetes status was significant (p-value < 2.2e-16), indicating that smoking is strongly associated with diabetes.

In the diabetic group, the percentage of those who actively engage in physical activity is 63.1%, compared with 77.6% in the non-diabetic group. The chi-square test for physical activity (PhysActivity) also showed a significant association with diabetes (p-value < 2.2e-16), suggesting that individuals actively engaging in physical activity are less likely to have diabetes.

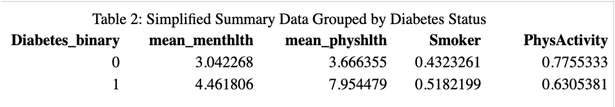

Table 2: Simplified Summary Data Grouped by Diabetes Status

Diabetes_binary	mean_menthlth	mean_physhlth	Smoker	PhysActivity
0	3.042268	3.666355	0.4323261	0.7755333
1	4.461806	7.954479	0.5182199	0.6305381

Correlation Analysis:

A Pearson correlation between poor mental health (MentHlth) and poor physical health (PhysHlth) revealed a moderate positive correlation (correlation coefficient = 0.38, p-value < 2.2e-16), indicating that these factors are moderately related.

B. Logistic Regression

The dataset was split into 50% training and 50% testing subsets. The logistic regression model was fitted to predict diabetes status (Diabetes_binary) using MentHlth, PhysHlth, Smoker, and PhysActivity.

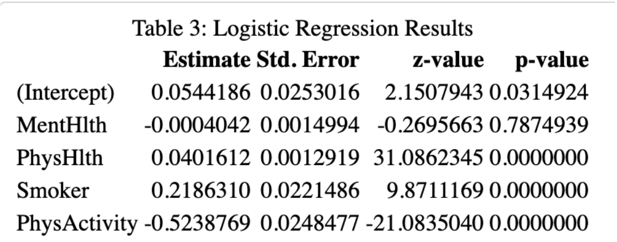

Table 3: Logistic Regression Results

	Estimate	Std. Error	z-value	p-value
(Intercept)	0.0544186	0.0253016	2.1507943	0.0314924
MentHlth	-0.0004042	0.0014994	-0.2695663	0.7874939
PhysHlth	0.0401612	0.0012919	31.0862345	0.0000000
Smoker	0.2186310	0.0221486	9.8711169	0.0000000
PhysActivity	-0.5238769	0.0248477	-21.0835040	0.0000000

For mental health, the estimate is pretty small, meaning that the log-odds of diabetes decrease very slightly for each additional day of poor mental health. And the p-value = 0.7875, which is not significant. This shows that mental health does not significantly contribute to predicting diabetes when adjusted for other factors.

For physical health, the estimate is 0.04, meaning that each additional day of poor physical health increases the log-odds of diabetes by 0.04. The Odds Ratio is 1.041, which means each additional day of poor physical health increases the odds of diabetes by around 4.1%. P-value is extremely small, showing the relationship is statistically significant.

In terms of smoking, being a smoker increases the log-odds of diabetes by 0.219 (Odds Ratio: 1.244), meaning that smokers have 24.4% higher odds of having diabetes compared to non-smokers. The small P-value shows that smoking is strongly associated with diabetes.

For physical activity, being physically active decreases the log-odds of diabetes by 0.524 (Odds Ratio: 0.592), representing that physically active individuals have 59.2% lower odds of having diabetes. The extremely small P-value shows that physical activity is a significantly strong predictive factor of diabetes.

C. Model Performance

The confusion matrix indicated moderate predictive performance, with an overall accuracy of 60.6%. The sensitivity (recall) was 51.5%, indicating that the model correctly identified 51.5% of diabetic individuals. This means that it missed many diabetic cases, limiting its ability to identify high-risk individuals. Specificity was higher at 69.6%, showing better performance in identifying non-diabetic individuals.

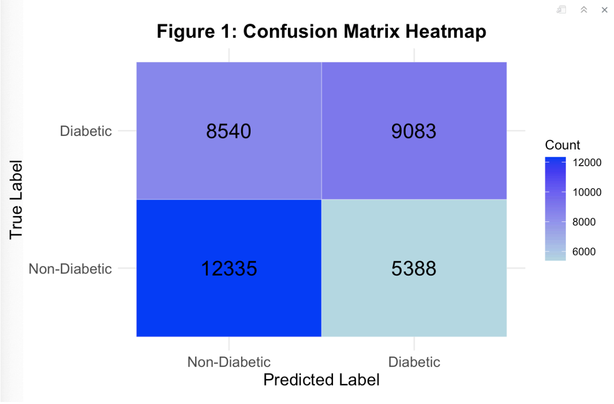

Figure 1: Confusion Matrix Heatmap

The ROC curve (Figure 2) yielded an Area Under the Curve (AUC) of 0.6468, suggesting fair predictive power of the logistic regression model in distinguishing between diabetic and non-diabetic individuals.

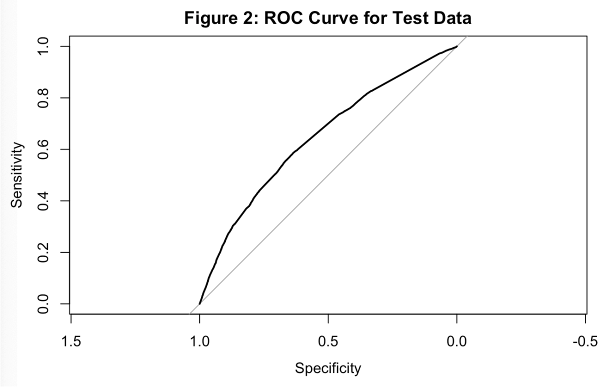

Figure 2: ROC Curve for Test Data

IV. CONCLUSIONS

This study aimed to explore the associations between health and lifestyle factors and diabetes status, as well as to evaluate the predictive power of a logistic regression model for identifying diabetes. The analysis revealed significant relationships between physical health, smoking, physical activity, and diabetes. Diabetic individuals were found to report more poor physical health days, with smoking positively and physical activity negatively associated with diabetes. The correlation analysis highlighted a moderate relationship between poor mental and physical health, emphasizing the interconnected nature of these factors in individuals with diabetes.

The logistic regression model demonstrated fair predictive performance, with an AUC of 0.6468, moderate accuracy, and higher specificity than sensitivity. While the model effectively identified non-diabetic individuals, its lower sensitivity indicates potential limitations in identifying diabetic individuals. These findings underscore the importance to address lifestyle factors such as smoking and physical activity to manage and prevent diabetes. Further research could incorporate additional variables, such as socioeconomic factors or genetic predispositions, or interactions to enhance predictive accuracy.

V. DICUSSION

Several limitations should be addressed to improve the reliability and applicability of the findings.

First, the logistic regression model demonstrates relatively low sensitivity, indicating that many diabetic cases were not correctly classified. To address this limitation, we could consider adjusting the classification threshold from the default value of 0.5 to a lower value, which may favor capturing more positive cases. Additionally, including more features or interaction terms in the model could better capture the variability among individuals with diabetes, potentially improving the model's sensitivity without significantly hurting overall performance. Features such as dietary habits, family history, or socioeconomic factors might provide additional predictive power.

Second, the role of mental health in diabetes needs further investigation. The Welch Two Sample t-test revealed a statistically significant difference in mental health between individuals with and without diabetes. However, in the logistic regression model, mental health did not seem like a significant predictor of diabetes, as indicated by its high p-value. This inconsistency may result from potential confounders and highlights the complexity of the relationship between mental health and diabetes, which may require further exploration.

Future research should explore these limitations by testing different model thresholds, adding additional features, and using better modeling techniques to improve the ability to identify diabetic cases and clarify the role of mental health in diabetes. Such efforts could lead to more comprehensive prevention and intervention tailored to individuals.

REFERENCES

1. World Health Organization. "Diabetes." *World Health Organization*, 14 Nov. 2024, www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 1 Dec. 2024.