Insights into the Relationship between Physical Activity and Diabetes Prevalence

Data to Paper

February 19, 2024

Abstract

Diabetes is a globally significant public health concern on the rise. The impact of physical activity on diabetes has been extensively studied, but there is still limited understanding of the specific relationship and the potential moderating effect of BMI. To address this research gap, we conducted an analysis using a comprehensive dataset extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) in 2015. This dataset includes diabetes-related factors obtained through a health-related telephone survey of over 400,000 Americans. By applying logistic regression models, we discovered a negative association between physical activity and diabetes prevalence, independent of confounders such as age, smoking status, and education level. Furthermore, our findings indicate that BMI moderates the association between physical activity and diabetes prevalence, with a heightened protective effect observed among individuals with higher BMI. These results underscore the significance of promoting physical activity as a preventive measure against diabetes, particularly among populations at higher risk. Consideration should be given to the limitations of self-reported data in our study. Nonetheless, our findings provide valuable insights for the development of targeted interventions and public health strategies to mitigate the burden of diabetes.

Introduction

Diabetes presents a growing global health challenge, with an increase in prevalence leading to significant public health and socio-economic implications [1, 2]. Obesity and physical inactivity are acknowledged as primary risk factors for the onset of the disease [3, 4, 5]. The role of physical activity as a preventative and management strategy for diabetes is well established [6, 7, 8].

While it is known that physical activity can lower diabetes prevalence [9, 10], our understanding is still limited on how this relationship is influenced by the Body Mass Index (BMI). What remains unclear is the extent to which BMI moderates the effect of physical activity on diabetes prevalence [7, 11].

To address these gaps, the current study leverages a comprehensive dataset from the Behavioral Risk Factor Surveillance System (BRFSS), a health-related survey conducted by the Centers for Disease Control and Prevention (CDC) in 2015 [12]. The vastness of this dataset, spanning 253,680 responses with 22 features including diabetes-related and demographic factors, offers a unique opportunity to delve deeper into the intricate association between physical activity, BMI, and diabetes prevalence. The BRFSS dataset has previously served as a robust resource for a multitude of health studies and offers great promise for expanding our understanding in this context [13].

Our research approach integrates logistic regression to model the relationship between diabetes prevalence and a set of explanatory variables including physical activity and other potential confounders [14]. This study presents significant insights into the negative association of physical activity with diabetes prevalence. Additionally, it reveals how this relationship is moderated by BMI, emphasizing the enhanced protective effect of physical activity for individuals with a higher BMI.

Results

In this study, we investigated the relationship between physical activity and diabetes prevalence using a comprehensive dataset extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) in 2015. With a total of 253,680 observations, our aim was to shed light on the association between physical activity and diabetes prevalence, adjusting for potential confounders, and to elucidate the potential moderating effect of BMI.

First, we conducted a detailed analysis of the distribution of physical activity stratified by diabetes status. As demonstrated in Table 1, individuals without diabetes seemed to engage in physical activity more frequently with an average of 0.777, compared to those with diabetes whose average was 0.631.

With the goal of examining the direct effect of physical activity on diabetes prevalence, we controlled for Age, BMI, Smoking status, High Blood Pressure, High Cholesterol, Education, and Income through logistic regression analysis. Results from Table 2 showed that physical activity had a

Table 1: Descriptive statistics of Physical Activity stratified by Diabetes

	Mean	Std.Dev
No Diabetes	0.777	0.416
Diabetes	0.631	0.483

significant negative association with diabetes, with a coefficient of -0.232 (SE=0.0136, p-value $< 10^{-6}$) which means a lower odd of having diabetes among individuals who engaged in physical activity. Our model revealed a pseudo R-squared of 0.1677, which indicated that the model able to explain a significant portion of the variation in diabetes incidence.

Table 2: Association between physical activity and diabetes prevalence

	Coef.	Std.Err.	Z	P-value	[0.025	0.975]
Intercept	-4.78	0.0526	-91	$<10^{-6}$	-4.89	-4.68
Physical Act.	-0.232	0.0136	-17.1	$<10^{-6}$	-0.258	-0.205
Age	0.132	0.00256	51.5	$<10^{-6}$	0.127	0.137
BMI	0.0696	0.000881	79	$<10^{-6}$	0.0679	0.0714
Smoker	0.0985	0.0126	7.84	$<10^{-6}$	0.0739	0.123
High BP	0.951	0.0144	66.1	$<10^{-6}$	0.923	0.979
High Chol	0.693	0.0132	52.4	$<10^{-6}$	0.667	0.719
Ed.	-0.0858	0.00676	-12.7	$<10^{-6}$	-0.0991	-0.0726
Income	-0.11	0.0032	-34.5	$<10^{-6}$	-0.117	-0.104

Physical Act.: 0: No activity, 1: Any activity
BMI: Body Mass Index
Age: Age in intervals of 5 years. 1: 18-24 –> 13: 80+ yrs
Smoker: 0: Non-smoker, 1: Smoker
High BP: 0: No high BP, 1: High BP
High Chol: 0: No high chol, 1: High chol
Ed.: Education level. 1: None –> 6: College
Income: Income category. 1: <=10K –> 8: >75K
Z: Standardized test statistic
P-value: Significance level of the Z statistic

Lastly, we sought to determine whether BMI has a moderating effect on the association between physical activity and diabetes. This was done by introducing an interaction term of Physical Activity and BMI to the logistic regression model. Interestingly, our results revealed a significant interaction term (coefficient =0.00906, SE=0.00174, p-value $<10^{-6}$) according to Table

3. These findings point out that the protective effect of physical activity against diabetes is stronger among individuals with higher BMI, and the model slightly improved with a pseudo R-squared of 0.1678.

Table 3: Moderating effect of BMI on the association between physical activity and diabetes

	Coef.	Std.Err.	Z	P-value	[0.025	0.975]
Intercept	-4.61	0.0617	-74.7	$<10^{-6}$	-4.73	-4.49
Physical Act.	-0.514	0.0558	-9.21	$<10^{-6}$	-0.624	-0.405
BMI	0.064	0.00138	46.3	$<10^{-6}$	0.0613	0.0667
Activity*BMI	0.00906	0.00174	5.21	$<10^{-6}$	0.00565	0.0125
Age	0.132	0.00256	51.6	$<10^{-6}$	0.127	0.137
Smoker	0.0977	0.0126	7.77	$<10^{-6}$	0.073	0.122
High BP	0.95	0.0144	66	$<10^{-6}$	0.922	0.978
High Chol	0.693	0.0132	52.4	$<10^{-6}$	0.667	0.719
Ed.	-0.0847	0.00676	-12.5	$<10^{-6}$	-0.098	-0.0715
Income	-0.11	0.0032	-34.4	$<10^{-6}$	-0.116	-0.104

Physical Act.: 0: No activity, 1: Any activity
BMI: Body Mass Index
Age: Age in intervals of 5 years. 1: 18-24 –> 13: 80+ yrs
Smoker: 0: Non-smoker, 1: Smoker
High BP: 0: No high BP, 1: High BP
High Chol: 0: No high chol, 1: High chol
Ed.: Education level. 1: None –> 6: College
Income: Income category. 1: <=10K –> 8: >75K
Z: Standardized test statistic
P-value: Significance level of the Z statistic
Activity*BMI: Interaction term between physical activity and BMI

In summary, our analysis reveals a significant independent negative association between physical activity and diabetes prevalence. Furthermore, BMI acts as a moderating variable in this relationship, thus accentuating the protective role of physical activity especially among individuals with higher BMI. These findings underscore the importance of promoting physical activity, particularly in high-risk populations, as a proactive measure against diabetes.

Discussion

Our study investigated the critical relationship between physical activity and the prevalence of diabetes, with a novel focus on the moderating effect of the Body Mass Index (BMI) [4, 7]. The motivation for this investigation emerged from the growing global incidence of diabetes and established risk factors, namely obesity and lack of physical activity [1, 2].

Employing the 2015 dataset from the Behavioral Risk Factor Surveillance System (BRFSS), we utilized logistic regression models to analyze the association while controlling for potential confounders. Crucial factors such as age, BMI, smoking status, high blood pressure, high cholesterol levels, education, and income were accounted for as these can confound the relationship between physical activity and diabetes prevalence [14].

By contrast with prior research, our results demonstrated a marked negative association between physical activity and diabetes prevalence, substantiating previous findings in the literature [9, 10]. Our study further innovates by revealing that BMI significantly strengthens the protective effect of physical activity against diabetes, thus heightening the contrast with previous studies that did not consider this interaction [7, 11].

However, our study has several limitations to consider. Particularly, the application of self-reported data introduces the potential for bias and inaccuracies. The self-reported nature of many factors in our study is a common issue in diabetes-related explorations[1]. Additionally, the cross-sectional design of our data prevents us from inferring causal relationships. The interplay between physical activity, diabetes prevalence, and BMI certainly has the potential to be influenced by other unmeasured factors.

Despite these limitations, our findings offer invaluable insights into the relationship between diabetes prevalence, physical activity, and BMI. For populations experiencing a sharp increase in diabetes, our results underscore the urgent need for promoting physical activity, especially among individuals in higher BMI categories.

For future work, a longitudinal study design can provide comprehensive insights into the relationship dynamics between physical activity, BMI, and diabetes prevalence over time [3, 5, 4]. Moreover, a deeper investigation into the role of different types of physical activity and potential mediators could be consequential in mitigating diabetes risk. Such studies would significantly contribute to the drafting of effective and inclusive public health strategies and policies.

Methods

Data Source

The dataset used in this study was obtained from the Behavioral Risk Factor Surveillance System (BRFSS) conducted by the Centers for Disease Control and Prevention (CDC) in the year 2015. The BRFSS is an annual health-related telephone survey that collects information on various health-related risk behaviors, chronic health conditions, and the use of preventative services from a large sample of Americans. The dataset used in this study consists of 253,680 responses with 22 features, including diabetes-related factors and demographic information.

Data Preprocessing

The dataset was preprocessed to ensure data quality. The original dataset was cleaned, removing any missing values. As a result, the final dataset used for analysis contained 253,680 responses with no missing values.

Data Analysis

To investigate the relationship between physical activity and the prevalence of diabetes, logistic regression analysis was conducted. Logistic regression is a statistical approach used to model the relationship between a binary dependent variable (in this case, diabetes prevalence) and one or more independent variables (including physical activity and other potential confounders).

Two separate logistic regression models were fitted to the dataset. These models were implemented using appropriate statistical software without specifying the specific version.

The first logistic regression model examined the association between physical activity and diabetes prevalence. This model included physical activity as the main independent variable, controlling for potential confounding factors such as age, smoking status, body mass index (BMI), high blood pressure, high cholesterol, education level, and income.

The second logistic regression model aimed to explore the moderating effect of BMI on the association between physical activity and diabetes prevalence. In addition to the variables included in the first model, this model included an interaction term between physical activity and BMI. This interaction term allows us to assess whether the relationship between physical

activity and diabetes prevalence differs for individuals with different BMI levels.

The logistic regression models were fitted using maximum likelihood estimation, which is a common method for estimating the parameters in logistic regression. Odds ratios (ORs) and corresponding 95% confidence intervals (CIs) were calculated to quantify the strength and significance of the associations. Pseudo R-squared values were also calculated to assess the goodness of fit of the models.

All analyses were performed using appropriate statistical software, and the resulting tables and additional results were saved for further use.

Code Availability

Custom code used to perform the data preprocessing and analysis, as well as the raw code outputs, are provided in Supplementary Methods.

A Data Description

Here is the data description, as provided by the user:

```
The dataset includes diabetes related factors extracted from the CDC's
    Behavioral Risk Factor Surveillance System (BRFSS), year 2015.
The original BRFSS, from which this dataset is derived, is a health-related
    telephone survey that is collected annually by the CDC.
Each year, the survey collects responses from over 400,000 Americans on health-
    related risk behaviors, chronic health conditions, and the use of preventative
    services. These features are either questions directly asked of participants, or
    calculated variables based on individual participant responses.


1 data file:

"diabetes_binary_health_indicators_BRFSS2015.csv"
The csv file is a clean dataset of 253,680 responses (rows) and 22 features
    (columns).
All rows with missing values were removed from the original dataset; the current
    file contains no missing values.

The columns in the dataset are:

#1 `Diabetes_binary`: (int, bool) Diabetes (0=no, 1=yes)
#2 `HighBP`: (int, bool) High Blood Pressure (0=no, 1=yes)
#3 `HighChol`: (int, bool) High Cholesterol (0=no, 1=yes)
#4 `CholCheck`: (int, bool) Cholesterol check in 5 years (0=no, 1=yes)
#5 `BMI`: (int, numerical) Body Mass Index
#6 `Smoker`: (int, bool) (0=no, 1=yes)
#7 `Stroke`: (int, bool) Stroke (0=no, 1=yes)
#8 `HeartDiseaseorAttack': (int, bool) coronary heart disease (CHD) or
    myocardial infarction (MI), (0=no, 1=yes)
#9 `PhysActivity`: (int, bool) Physical Activity in past 30 days (0=no, 1=yes)
#10 `Fruits`: (int, bool) Consume one fruit or more each day (0=no, 1=yes)
#11 `Veggies`: (int, bool) Consume one Vegetable or more each day (0=no, 1=yes)
#12 `HvyAlcoholConsump` (int, bool) Heavy drinkers (0=no, 1=yes)
#13 `AnyHealthcare` (int, bool) Have any kind of health care coverage (0=no,
    1=yes)
#14 `NoDocbcCost` (int, bool) Was there a time in the past 12 months when you
```

```
    needed to see a doctor but could not because of cost? (0=no, 1=yes)
#15 `GenHlth` (int, ordinal) self-reported health (1=excellent, 2=very good,
    3=good, 4=fair, 5=poor)
#16 `MentHlth` (int, ordinal) How many days during the past 30 days was your
    mental health not good? (1-30 days)
#17 `PhysHlth` (int, ordinal) Hor how many days during the past 30 days was your
    physical health not good? (1-30 days)
#18 `DiffWalk` (int, bool) Do you have serious difficulty walking or climbing
    stairs? (0=no, 1=yes)
#19 `Sex` (int, categorical) Sex (0=female, 1=male)
#20 `Age` (int, ordinal) Age, 13-level age category in intervals of 5 years
    (1=18-24, 2=25-29, ..., 12=75-79, 13=80 or older)
#21 `Education` (int, ordinal) Education level on a scale of 1-6 (1=Never
    attended school, 2=Elementary, 3=Some high school, 4=High school, 5=Some
    college, 6=College)
#22 `Income` (int, ordinal) Income scale on a scale of 1-8 (1=<=10K, 2=<=15K,
    3=<=20K, 4=<=25K, 5=<=35K, 6=<=50K, 7=<=75K, 8=>75K)
```

B Data Exploration

B.1 Code

The Data Exploration was carried out using the following custom code:

```

import pandas as pd
import numpy as np

# Load the data
df =
    ↪ pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.csv")

# Open the output file
with open("data_exploration.txt", "w") as out_file:

    # Data Size
    out_file.write("# Data Size\n")
    out_file.write(f"Number of Rows: {df.shape[0]}\n")
```

```
    out_file.write(f"Number of Columns: {df.shape[1]}\n")
    out_file.write("\n")

    # Summary Statistics
    out_file.write("# Summary Statistics\n")
    summary_stats = df.describe().transpose()
    summary_stats.to_string(out_file)
    out_file.write("\n")

    # Categorical Variables
    out_file.write("# Categorical Variables\n")
    categorical_vars =
    ↪  df.select_dtypes(include=['object']).columns
    for var in categorical_vars:
        out_file.write(f"\nFor variable {var}, the top 5 most
        ↪  common values are:\n")
        freq = df[var].value_counts().head(5).to_string()
        out_file.write(f"{freq}\n")
    out_file.write("\n")

    # Missing Values
    out_file.write("# Missing Values\n")
    missing_values = df.isnull().sum()
    missing_percentage = missing_values / df.shape[0] * 100
    out_file.write("Number and Percentage of Missing Values
    ↪  for Each Variable are as follows:\n")
    for (col, missing), percentage in
    ↪  zip(missing_values.items(), missing_percentage):
        out_file.write(f"{col}: Count = {missing}, Percentage
        ↪  = {percentage}%\n")
    out_file.write("\n")

    # Balance of Target Variable (Diabetes_binary)
    out_file.write("# Balance of Target Variable
    ↪  (Diabetes_binary)\n")
    target_balance = df['Diabetes_binary'].value_counts()
    target_percent =
    ↪  df['Diabetes_binary'].value_counts(normalize=True) *
    ↪  100
```

```
    out_file.write("Count and Percentage for Each Class of Target Variable are as follows:\n")
    for (val, balance), percent in zip(target_balance.items(), target_percent):
        out_file.write(f"Class {val}: Count = {balance}, Percentage = {percent}%\n")

# Close the output file
out_file.close()
```

B.2 Code Description

The code performs data exploration on a dataset containing diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The goal is to gain insights into the dataset and understand its characteristics.

The code begins by loading the dataset into a pandas DataFrame. It then proceeds to perform several analysis steps:

1. Data Size: The code obtains the number of rows and columns in the dataset and writes this information to the output file.

2. Summary Statistics: The code calculates summary statistics for each numerical variable in the dataset, including the count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum. It writes these statistics to the output file.

3. Categorical Variables: The code identifies the categorical variables in the dataset and finds the top 5 most common values for each of these variables. It writes these values to the output file.

4. Missing Values: The code identifies any missing values in the dataset by counting the number of missing values for each variable. It also calculates the percentage of missing values for each variable. The code then writes the count and percentage of missing values for each variable to the output file.

5. Balance of Target Variable (Diabetes_binary): The code determines the balance of the target variable, "Diabetes_binary," which indicates whether a person has diabetes or not. It calculates the count and percentage of each class (0=no diabetes, 1=diabetes) and writes this information to the output file.

Finally, the code writes the output to a text file named "data_exploration.txt". The file contains the results of the data exploration analysis, including the data size, summary statistics, categorical variables, missing values, and the

balance of the target variable.

The information provided in the "data_exploration.txt" file can be used to understand the structure and distribution of the dataset, identify any data quality issues such as missing values, and gain initial insights for further analysis.

B.3 Code Output

data_exploration.txt

```
# Data Size
Number of Rows: 253680
Number of Columns: 22

# Summary Statistics
                       count     mean     std  min  25%  50%  75%  max
Diabetes_binary       253680   0.1393  0.3463    0    0    0    0    1
HighBP                253680    0.429  0.4949    0    0    0    1    1
HighChol              253680   0.4241  0.4942    0    0    0    1    1
CholCheck             253680   0.9627  0.1896    0    1    1    1    1
BMI                   253680    28.38   6.609   12   24   27   31   98
Smoker                253680   0.4432  0.4968    0    0    0    1    1
Stroke                253680  0.04057  0.1973    0    0    0    0    1
HeartDiseaseorAttack  253680  0.09419  0.2921    0    0    0    0    1
PhysActivity          253680   0.7565  0.4292    0    1    1    1    1
Fruits                253680   0.6343  0.4816    0    0    1    1    1
Veggies               253680   0.8114  0.3912    0    1    1    1    1
HvyAlcoholConsump     253680   0.0562  0.2303    0    0    0    0    1
AnyHealthcare         253680   0.9511  0.2158    0    1    1    1    1
NoDocbcCost           253680  0.08418  0.2777    0    0    0    0    1
GenHlth               253680    2.511   1.068    1    2    2    3    5
MentHlth              253680    3.185   7.413    0    0    0    2   30
PhysHlth              253680    4.242   8.718    0    0    0    3   30
DiffWalk              253680   0.1682  0.3741    0    0    0    0    1
Sex                   253680   0.4403  0.4964    0    0    0    1    1
Age                   253680    8.032   3.054    1    6    8   10   13
Education             253680     5.05  0.9858    1    4    5    6    6
Income                253680    6.054   2.071    1    5    7    8    8
# Categorical Variables
```

```
# Missing Values
Number and Percentage of Missing Values for Each Variable are as follows:
Diabetes_binary: Count = 0, Percentage = 0.0%
HighBP: Count = 0, Percentage = 0.0%
HighChol: Count = 0, Percentage = 0.0%
CholCheck: Count = 0, Percentage = 0.0%
BMI: Count = 0, Percentage = 0.0%
Smoker: Count = 0, Percentage = 0.0%
Stroke: Count = 0, Percentage = 0.0%
HeartDiseaseorAttack: Count = 0, Percentage = 0.0%
PhysActivity: Count = 0, Percentage = 0.0%
Fruits: Count = 0, Percentage = 0.0%
Veggies: Count = 0, Percentage = 0.0%
HvyAlcoholConsump: Count = 0, Percentage = 0.0%
AnyHealthcare: Count = 0, Percentage = 0.0%
NoDocbcCost: Count = 0, Percentage = 0.0%
GenHlth: Count = 0, Percentage = 0.0%
MentHlth: Count = 0, Percentage = 0.0%
PhysHlth: Count = 0, Percentage = 0.0%
DiffWalk: Count = 0, Percentage = 0.0%
Sex: Count = 0, Percentage = 0.0%
Age: Count = 0, Percentage = 0.0%
Education: Count = 0, Percentage = 0.0%
Income: Count = 0, Percentage = 0.0%

# Balance of Target Variable (Diabetes_binary)
Count and Percentage for Each Class of Target Variable are as follows:
Class 0: Count = 218334, Percentage = 86.07           %
Class 1: Count = 35346, Percentage = 13.93            %
```

C Data Analysis

C.1 Code

The Data Analysis was carried out using the following custom code:

```

# IMPORT
import pandas as pd
```

```python
import numpy as np
import pickle
import statsmodels.formula.api as smf

# LOAD DATA
df = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.csv")

# DATASET PREPARATIONS
# No dataset preparations are needed as there are no missing values.
# Nor does the dataset require any aggregated operations or dataset joining operations.

# DESCRIPTIVE STATISTICS
## Table 0: "Descriptive statistics of Physical Activity stratified by Diabetes"
df0 = df.groupby('Diabetes_binary')['PhysActivity'].agg(['mean', 'std'])
df0.index = ['No Diabetes', 'Diabetes']
df0.to_pickle('table_0.pkl')

# PREPROCESSING
# No preprocessing is needed as data is already cleaned. All variables are either binary or ordinal and there're no categorical variables that require dummy variables.
# The numerical features do not require normalization or standardization as they will be fed to a logistic regression model that is agnostic to the scale of the features.

# ANALYSIS
## Table 1: "Association between physical activity and diabetes prevalence"
formula = "Diabetes_binary ~ PhysActivity + Age + BMI + Smoker + HighBP + HighChol + Education + Income"
model1 = smf.logit(formula, data=df).fit()
df1 = model1.summary2().tables[1]
df1.to_pickle('table_1.pkl')
```

```
31
32 ## Table 2: "Moderating effect of BMI on the association
   ↪ between physical activity and diabetes prevalence"
33 formula = "Diabetes_binary ~ PhysActivity*BMI + Age + BMI +
   ↪ Smoker + HighBP + HighChol + Education + Income"
34 model2 = smf.logit(formula, data=df).fit()
35 df2 = model2.summary2().tables[1]
36 df2.to_pickle('table_2.pkl')
37
38 # SAVE ADDITIONAL RESULTS
39 additional_results = {
40  'Total number of observations': df.shape[0],
41  'Pseudo R-squ of model1': model1.prsquared,
42  'Pseudo R-squ of model2': model2.prsquared,
43 }
44
45 with open('additional_results.pkl', 'wb') as f:
46  pickle.dump(additional_results, f)
47
```

C.2 Code Description

The code performs data analysis on a dataset of diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015.

First, the code loads the dataset into a pandas DataFrame. The dataset is already clean and contains no missing values.

Descriptive statistics are then calculated for the variable "Physical Activity" stratified by the presence of diabetes. The mean and standard deviation of physical activity are calculated for both the group with diabetes and the group without diabetes.

Next, the code prepares the data for analysis. Since the dataset is already clean, no additional preprocessing is required. The variables in the dataset are either binary or ordinal, and there are no categorical variables that need to be transformed into dummy variables. The numerical features do not require normalization or standardization as they will be used in a logistic regression model, which is scale-agnostic.

The code then performs two separate analyses using logistic regression models.

In the first analysis, the code examines the association between physical

activity and diabetes prevalence while controlling for other variables such as age, body mass index (BMI), smoking status, high blood pressure, high cholesterol, education, and income. The logistic regression model is fitted using the formula "Diabetes_binary ~ PhysActivity + Age + BMI + Smoker + HighBP + HighChol + Education + Income". The results of this analysis, including the coefficients and p-values for each variable, are saved in a DataFrame.

In the second analysis, the code investigates whether the association between physical activity and diabetes prevalence is moderated by BMI. This is done by including an interaction term between physical activity and BMI in the logistic regression model. The formula for this model is "Diabetes_binary ~ PhysActivity*BMI + Age + BMI + Smoker + HighBP + HighChol + Education + Income". The results of this analysis are also saved in a DataFrame.

Finally, the code saves additional results in a pickle file named "additional_results.pkl". These results include the total number of observations in the dataset, as well as the pseudo R-squared values for both logistic regression models.

Overall, the code performs exploratory analysis and hypothesis testing to examine the associations between physical activity, diabetes prevalence, and other relevant variables in the dataset. The logistic regression models provide insights into the relationships and potential moderating effects, while the additional results provide summary statistics and model evaluation metrics.

C.3 Code Output

table_0.pkl

	mean	std
No Diabetes	0.7769	0.4163
Diabetes	0.6305	0.4827

table_1.pkl

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-4.783	0.05257	-90.99	0	-4.886	-4.68
PhysActivity	-0.2318	0.01358	-17.07	2.537e-65	-0.2584	-0.2052
Age	0.132	0.002561	51.53	0	0.1269	0.137
BMI	0.06963	0.0008814	79	0	0.0679	0.07135
Smoker	0.09853	0.01257	7.837	4.62e-15	0.07389	0.1232

HighBP	0.9511	0.01439	66.07	0	0.9229	0.9793
HighChol	0.6933	0.01322	52.42	0	0.6674	0.7192
Education	-0.08583	0.006756	-12.7	5.562e-37	-0.09907	-0.07259
Income	-0.1103	0.0032	-34.48	1.672e-260	-0.1166	-0.104

table_2.pkl

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-4.612	0.06173	-74.72	0	-4.733	-4.491
PhysActivity	-0.5142	0.05584	-9.209	3.307e-20	-0.6236	-0.4047
BMI	0.06402	0.001384	46.27	0	0.06131	0.06674
PhysActivity:BMI	0.009064	0.00174	5.209	1.897e-07	0.005654	0.01247
Age	0.1323	0.002562	51.63	0	0.1272	0.1373
Smoker	0.09765	0.01257	7.767	8.023e-15	0.07301	0.1223
HighBP	0.9498	0.0144	65.97	0	0.9216	0.978
HighChol	0.6931	0.01322	52.42	0	0.6672	0.719
Education	-0.08471	0.006757	-12.54	4.677e-36	-0.09796	-0.07147
Income	-0.1102	0.003199	-34.44	5.886e-260	-0.1164	-0.1039

additional_results.pkl


```
{
    'Total number of observations': 253680,
    'Pseudo R-squ of model1': 0.1677             ,
    'Pseudo R-squ of model2': 0.1678            ,
}
```


D LaTeX Table Design

D.1 Code

The LaTeX Table Design was carried out using the following custom code:

```

# IMPORT
import pandas as pd
from typing import Dict, Tuple, Optional
from my_utils import to_latex_with_note, format_p_value

Mapping = Dict[str, Tuple[Optional[str], Optional[str]]]

```

```python

# PREPARATION FOR ALL TABLES
def split_mapping(d: Mapping):
    abbrs_to_names = {abbr: name for abbr, (name, definition) in d.items() if name is not None}
    names_to_definitions = {name or abbr: definition for abbr, (name, definition) in d.items() if definition is not None}
    return abbrs_to_names, names_to_definitions

shared_mapping: Mapping = {
    'PhysActivity': ('Physical Act.', '0: No activity, 1: Any activity'),
    'BMI': ('BMI', "Body Mass Index"),
    'Age': ('Age', 'Age in intervals of 5 years. 1: 18-24 --> 13: 80+ yrs'),
    'Smoker': ('Smoker', '0: Non-smoker, 1: Smoker'),
    'HighBP': ('High BP', '0: No high BP, 1: High BP'),
    'HighChol': ('High Chol', '0: No high chol, 1: High chol'),
    'Education': ('Ed.', 'Education level. 1: None --> 6: College'),
    'Income': ('Income', 'Income category. 1: <=10K --> 8: >75K'),
    'Coef.': ('Coef.', None),
    'Std.Err.': ('Std.Err.', None),
    'z': ('Z', 'Standardized test statistic'),
    'P>|z|': ('P-value', 'Significance level of the Z statistic')
}

# TABLE 0:
df0 = pd.read_pickle('table_0.pkl')

mapping0 = {'mean': ('Mean', None), 'std': ('Std.Dev', None)}
abbrs_to_names, legend = split_mapping(mapping0)
df0.rename(columns=abbrs_to_names, index=abbrs_to_names, inplace=True)

to_latex_with_note(
```

```
    df0,
    'table_0.tex',
    caption="Descriptive statistics of Physical Activity stratified by Diabetes",
    label='table:descriptive_statistics_0',
    legend=legend)

# TABLE 1:
df1 = pd.read_pickle('table_1.pkl')
abbrs_to_names, legend = split_mapping(shared_mapping)
df1.rename(columns=abbrs_to_names, index=abbrs_to_names, inplace=True)

# FORMAT VALUES
df1['P-value'] = df1['P-value'].apply(format_p_value)

to_latex_with_note(
    df1,
    'table_1.tex',
    caption="Association between  physical activity and diabetes prevalence",
    label='table:activity_and_diabetes_1',
    legend=legend)

# TABLE 2:
df2 = pd.read_pickle('table_2.pkl')

mapping2 = dict(shared_mapping.items())
mapping2.update({
    'PhysActivity:BMI': ('Activity*BMI', "Interaction term between physical activity and BMI"),
})

abbrs_to_names, legend = split_mapping(mapping2)
df2.rename(columns=abbrs_to_names, index=abbrs_to_names, inplace=True)

# FORMAT VALUES
df2['P-value'] = df2['P-value'].apply(format_p_value)

```

```
to_latex_with_note(
    df2,
    'table_2.tex',
    caption="Moderating effect of BMI on the association
    ↪  between physical activity and diabetes",
    label='table:activity_and_diabetes_2',
    legend=legend)

```

D.2 Code Output

table_0.tex

```
\begin{table}[h]
\caption{Descriptive statistics of Physical Activity stratified by Diabetes}
\label{table:descriptive_statistics_0}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrr}
\toprule
 & Mean & Std.Dev \\
\midrule
\textbf{No Diabetes} & 0.777 & 0.416 \\
\textbf{Diabetes} & 0.631 & 0.483 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
\begin{table}[h]
\caption{Association between  physical activity and diabetes prevalence}
\label{table:activity_and_diabetes_1}
\begin{threeparttable}
```

```
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err. & Z & P-value & [0.025 & 0.975] \\
\midrule
\textbf{Intercept} & -4.78 & 0.0526 & -91 & $<$1e-06 & -4.89 & -4.68 \\
\textbf{Physical Act.} & -0.232 & 0.0136 & -17.1 & $<$1e-06 & -0.258 & -0.205 \\
\textbf{Age} & 0.132 & 0.00256 & 51.5 & $<$1e-06 & 0.127 & 0.137 \\
\textbf{BMI} & 0.0696 & 0.000881 & 79 & $<$1e-06 & 0.0679 & 0.0714 \\
\textbf{Smoker} & 0.0985 & 0.0126 & 7.84 & $<$1e-06 & 0.0739 & 0.123 \\
\textbf{High BP} & 0.951 & 0.0144 & 66.1 & $<$1e-06 & 0.923 & 0.979 \\
\textbf{High Chol} & 0.693 & 0.0132 & 52.4 & $<$1e-06 & 0.667 & 0.719 \\
\textbf{Ed.} & -0.0858 & 0.00676 & -12.7 & $<$1e-06 & -0.0991 & -0.0726 \\
\textbf{Income} & -0.11 & 0.0032 & -34.5 & $<$1e-06 & -0.117 & -0.104 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Physical Act.}: 0: No activity, 1: Any activity
\item \textbf{BMI}: Body Mass Index
\item \textbf{Age}: Age in intervals of 5 years. 1: 18-24 --$>$ 13: 80+ yrs
\item \textbf{Smoker}: 0: Non-smoker, 1: Smoker
\item \textbf{High BP}: 0: No high BP, 1: High BP
\item \textbf{High Chol}: 0: No high chol, 1: High chol
\item \textbf{Ed.}: Education level. 1: None --$>$ 6: College
\item \textbf{Income}: Income category. 1: $<$=10K --$>$ 8: $>$75K
\item \textbf{Z}: Standardized test statistic
\item \textbf{P-value}: Significance level of the Z statistic
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_2.tex

```
\begin{table}[h]
\caption{Moderating effect of BMI on the association between physical activity
    and diabetes}
\label{table:activity_and_diabetes_2}
```

```
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err. & Z & P-value & [0.025 & 0.975] \\
\midrule
\textbf{Intercept} & -4.61 & 0.0617 & -74.7 & $<$1e-06 & -4.73 & -4.49 \\
\textbf{Physical Act.} & -0.514 & 0.0558 & -9.21 & $<$1e-06 & -0.624 & -0.405 \\
\textbf{BMI} & 0.064 & 0.00138 & 46.3 & $<$1e-06 & 0.0613 & 0.0667 \\
\textbf{Activity*BMI} & 0.00906 & 0.00174 & 5.21 & $<$1e-06 & 0.00565 & 0.0125
    \\
\textbf{Age} & 0.132 & 0.00256 & 51.6 & $<$1e-06 & 0.127 & 0.137 \\
\textbf{Smoker} & 0.0977 & 0.0126 & 7.77 & $<$1e-06 & 0.073 & 0.122 \\
\textbf{High BP} & 0.95 & 0.0144 & 66 & $<$1e-06 & 0.922 & 0.978 \\
\textbf{High Chol} & 0.693 & 0.0132 & 52.4 & $<$1e-06 & 0.667 & 0.719 \\
\textbf{Ed.} & -0.0847 & 0.00676 & -12.5 & $<$1e-06 & -0.098 & -0.0715 \\
\textbf{Income} & -0.11 & 0.0032 & -34.4 & $<$1e-06 & -0.116 & -0.104 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Physical Act.}: 0: No activity, 1: Any activity
\item \textbf{BMI}: Body Mass Index
\item \textbf{Age}: Age in intervals of 5 years. 1: 18-24 --$>$ 13: 80+ yrs
\item \textbf{Smoker}: 0: Non-smoker, 1: Smoker
\item \textbf{High BP}: 0: No high BP, 1: High BP
\item \textbf{High Chol}: 0: No high chol, 1: High chol
\item \textbf{Ed.}: Education level. 1: None --$>$ 6: College
\item \textbf{Income}: Income category. 1: $<$=10K --$>$ 8: $>$75K
\item \textbf{Z}: Standardized test statistic
\item \textbf{P-value}: Significance level of the Z statistic
\item \textbf{Activity*BMI}: Interaction term between physical activity and BMI
\end{tablenotes}
\end{threeparttable}
\end{table}
```

References

[1] S. Akter, Md. Mizanur Rahman, Sarah Krull Abe, and P. Sultana. Prevalence of diabetes and prediabetes and their risk factors among bangladeshi adults: a nationwide survey. *Bulletin of the World Health Organization*, 92 3:204–13, 213A, 2014.

[2] David W Lam and D. Leroith. The worldwide diabetes epidemic. *Current Opinion in Endocrinology & Diabetes and Obesity*, 19:9396, 2012.

[3] J. Chan, E. Rimm, G. Colditz, M. Stampfer, and W. Willett. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. *Diabetes Care*, 17:961 – 969, 1994.

[4] S. Wild and C. Byrne. Risk factors for diabetes and coronary heart disease. *BMJ : British Medical Journal*, 333:1009 – 1011, 2006.

[5] Gitanjali M Singh, G. Danaei, F. Farzadfar, G. Stevens, M. Woodward, D. Wormser, S. Kaptoge, G. Whitlock, Q. Qiao, S. Lewington, E. Di Angelantonio, S. Vander Hoorn, C. Lawes, Mohammed K. Ali, D. Mozaffarian, and M. Ezzati. The age-specific quantitative effects of metabolic risk factors on cardiovascular diseases and diabetes: A pooled analysis. *PLoS ONE*, 8, 2013.

[6] K. Fox. The influence of physical activity on mental well-being. *Public Health Nutrition*, 2:411 – 418, 1999.

[7] B. Bohn, A. Herbst, M. Pfeifer, D. Krakow, S. Zimny, F. Kopp, A. Melmer, J. Steinacker, and R. Holl. Impact of physical activity on glycemic control and prevalence of cardiovascular risk factors in adults with type 1 diabetes: A cross-sectional multicenter study of 18,028 patients. *Diabetes Care*, 38:1536 – 1543, 2015.

[8] K. Eckert. Impact of physical activity and bodyweight on health-related quality of life in people with type 2 diabetes. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 5:303 – 311, 2012.

[9] Fanfan Zhao, Wentao Wu, Xiaojie Feng, Chengzhuo Li, Didi Han, Xiaojuan Guo, and Jun Lyu. Physical activity levels and diabetes prevalence in us adults: Findings from nhanes 20152016. *Diabetes Therapy*, 11:1303 – 1316, 2020.

[10] Liang Shi, X. Shu, Hong-Lan Li, H. Cai, Qiaolan Liu, W. Zheng, Y. Xiang, and R. Villegas. Physical activity, smoking, and alcohol consumption in association with incidence of type 2 diabetes among middle-aged and elderly chinese men. *PLoS ONE*, 8, 2013.

[11] X. Pi-Sunyer, G. Blackburn, F. Brancati, G. Bray, R. Bright, J. Clark, J. M. Curtis, M. Espeland, J. Foreyt, Kathryn Graves, S. Haffner, B. Harrison, James O Hill, E. Horton, J. Jakicic, R. Jeffery, K. Johnson, S. Kahn, D. Kelley, A. Kitabchi, W. Knowler, C. Lewis, B. J. Maschak-Carey, B. Montgomery, D. Nathan, Jennifer Patricio, A. Peters, J. Redmon, R. Reeves, D. Ryan, M. Safford, B. van Dorsten, T. Wadden, L. Wagenknecht, Jacqueline Wesche-Thobaben, R. Wing, and S. Yanovski. Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes. *Diabetes Care*, 30:1374 – 1383, 2007.

[12] Lenzetta Rolle-Lake and E. Robbins. Behavioral risk factor surveillance system (brfss). 2020.

[13] V. Preedy and R. Watson. Behavioral risk factor surveillance system. *Iowa medicine : journal of the Iowa Medical Society*, 79 9:436, 438, 1989.

[14] S. Menard. Applied logistic regression analysis. 1996.