Insights into the Association between Physical Activity and Chronic Health Conditions in Individuals with Diabetes

Data to Paper

September 28, 2023

Abstract

Diabetes is a prevalent chronic health condition with significant public health implications. However, the relationship between physical activity and associated chronic health conditions in individuals with diabetes remains poorly understood. This study aims to address this research gap by analyzing a comprehensive dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS). Using logistic regression models, we examined the association between physical activity and three major chronic health conditions in individuals with diabetes: high blood pressure, high cholesterol, and coronary heart disease. Our findings reveal significant negative associations between physical activity and these chronic health conditions, even after adjusting for key factors such as age, sex, BMI, and smoking status. These results highlight the potential benefits of increasing physical activity levels for managing diabetes-related health concerns. Nonetheless, it is important to note that the accuracy of the statistical models employed was relatively low, likely due to the use of self-reported data. The findings of this study contribute to our understanding of the role of physical activity in the context of diabetes and provide valuable insights for healthcare professionals and policymakers in developing interventions to promote physical activity and improve health outcomes in individuals with diabetes.

Introduction

Diabetes, a major chronic health condition, is impacting a rapidly growing number of individuals worldwide [1]. This global burden is further aggravated by common comorbidities such as high blood pressure, high cholesterol, and coronary heart disease in diabetic individuals [2, 3, 4, 5]. Accumulating research suggests the merits of physical activity in promoting better health outcomes [6]. For individuals with diabetes, evidence corroborates the potential of physical activity in the management of concurrent chronic conditions [7, 8, 9], however, a nuanced understanding of how physical activity interacts with specific chronic health conditions among individuals with diabetes is lacking.

Considering the existing literature, physical activity has been shown to offer protective effects against adverse health outcomes [10, 11], a decrease in the risk of mortality in individuals with clustered metabolic risk factors [12, 13, 14], and a potential to constructively affect cardiovascular diseases [15, 16]. Importantly, studies like [3] and [4] especially underscore the positive influence of physical activity in the context of chronic diseases, including diabetes. Nevertheless, detailed research into associations between physical activity and major chronic health conditions, particularly high blood pressure, high cholesterol, and coronary heart disease, among individuals with diabetes, is still limited.

In an effort to contribute towards filling this research gap, the present study uses the 2015 Behavioral Risk Factor Surveillance System dataset [17]. This dataset, which has been extensively used in related studies [18, 19, 20, 21, 22, 23], presents an ideal platform to further examine the relationship between physical activity and the stated chronic conditions in the context of diabetes.

Adopting logistic regression models [24, 18, 25, 26, 27, 28], our analysis investigates the explicit associations between physical activity and each of the three chronic health conditions. The analysis accommodates potential confounding factors including age, sex, body mass index, and smoking status. Our findings underscore the potential favorability of physical activity for improved health outcomes in diabetes management, which are highlighted in ensuing sections.

Results

To understand the relationship between physical activity and chronic health conditions in individuals with diabetes, we conducted logistic regression analyses, adjusting for key factors such as age, sex, BMI, and smoking status.

First, we compared the prevalence of chronic health conditions between individuals with and without diabetes. As shown in Table 1, individuals with diabetes had a higher prevalence of high blood pressure (37.7% vs 75.3%), high cholesterol (38.4% vs 67%), and coronary heart disease or heart attack (7.34% vs 22.3%) compared to those without diabetes.

Table 1: Descriptive Statistics of Physical Activity and Chronic Health Conditions for both Diabetes and Non-Diabetes Individuals

	Phys. Act.	High BP	High Chol.	Heart Dis./Att.
Diabetes_binary				
No Diabetes	0.777	0.377	0.384	0.0734
Diabetes	0.631	0.753	0.67	0.223

Values represent the proportions of individuals
Phys. Act.: Physical Activity in past 30 days (0=no, 1=yes)
High BP: High Blood Pressure (0=no, 1=yes)
High Chol.: High Cholesterol (0=no, 1=yes)
Heart Dis./Att.: Coronary heart disease (CHD) or myocardial infarction (MI), (0=no, 1=yes)

Next, focusing on individuals with diabetes, we examined the association between physical activity and high blood pressure. Our logistic regression analysis (Table 2) revealed a negative association between physical activity and high blood pressure (coefficient = -0.172, SE = 0.0272, p-value $< 10^{-6}$). After adjusting for age, sex, BMI, and smoking status, individuals with diabetes who engaged in physical activity had a lower likelihood of having high blood pressure.

Table 2: Association between Physical Activity and High BP in Individuals with Diabetes

	Coeff.	Std Err.	z-score	P-value	CI Lower	CI Upper
Phys. Act.	-0.172	0.0272	-6.32	$<10^{-6}$	-0.225	-0.119

Values represent logistic regression coefficients. P-values are two-sided.
Phys. Act.: Physical Activity in past 30 days (0=no, 1=yes)
P-value: P-value of the logistic regression model
z-score: Z-score for the coefficient in the logistic regression model
Coeff.: Estimated model coefficient
Std Err.: Standard error for the estimated coefficient
CI Lower: 95% Confidence Interval Lower Bound
CI Upper: 95% Confidence Interval Upper Bound

Further, we investigated the association between physical activity and high cholesterol in individuals with diabetes. The logistic regression analysis (Table 3) showed a negative association between physical activity and high cholesterol (coefficient = -0.117, SE = 0.0241, p-value = 1.1 10^{-6}), even after adjusting for age, sex, BMI, and smoking status. The odds ratio of 0.8896 (95% CI: [0.8477, 0.9335]) indicated that individuals with diabetes who engaged in physical activity had a lower likelihood of having high cholesterol.

Table 3: Association between Physical Activity and High Chol. in Individuals with Diabetes

	Coeff.	Std Err.	z-score	P-value	CI Lower	CI Upper
Phys. Act.	-0.117	0.0241	-4.87	1.1 10^{-6}	-0.165	-0.0702

Values represent logistic regression coefficients. P-values are two-sided.
Phys. Act.: Physical Activity in past 30 days (0=no, 1=yes)
P-value: P-value of the logistic regression model
z-score: Z-score for the coefficient in the logistic regression model
Coeff.: Estimated model coefficient
Std Err.: Standard error for the estimated coefficient
CI Lower: 95% Confidence Interval Lower Bound
CI Upper: 95% Confidence Interval Upper Bound

Finally, we explored the association between physical activity and coronary heart disease in individuals with diabetes using logistic regression analysis (Table 4). After adjusting for age, sex, BMI, and smoking status, we found a significant negative association between physical activity and coronary heart disease (coefficient = -0.308, SE = 0.0272, p-value $< 10^{-6}$). This indicates that individuals with diabetes who engaged in physical activity had a lower likelihood of having coronary heart disease.

Table 4: Association between Physical Activity and Heart Dis./Att. in Individuals with Diabetes

	Coeff.	Std Err.	z-score	P-value	CI Lower	CI Upper
Phys. Act.	-0.308	0.0272	-11.3	$<10^{-6}$	-0.361	-0.255

Values represent logistic regression coefficients. P-values are two-sided.
Phys. Act.: Physical Activity in past 30 days (0=no, 1=yes)
P-value: P-value of the logistic regression model
z-score: Z-score for the coefficient in the logistic regression model
Coeff.: Estimated model coefficient
Std Err.: Standard error for the estimated coefficient
CI Lower: 95% Confidence Interval Lower Bound
CI Upper: 95% Confidence Interval Upper Bound

In summary, our logistic regression analyses demonstrated that physical activity is negatively associated with high blood pressure, high cholesterol, and coronary heart disease in individuals with diabetes, even after adjusting for age, sex, BMI, and smoking status. These findings suggest that increasing physical activity levels may have benefits for managing diabetes-related health concerns.

Discussion

At the backdrop of the escalating global diabetes epidemic [1], this study aimed to explore the role of physical activity, a cost-effective and accessible intervention, in managing diabetes-related health problems, specifically high blood pressure, high cholesterol, and coronary heart disease. These conditions, amongst the most prevalent comorbidities in individuals with diabetes, are of immediate public health concern [2, 3].

Employing logistic regression models and adjusting for key factors such as age, sex, BMI and smoking status, the study established a significant negative association between physical activity and the three chronic health conditions in question, amongst individuals with diabetes [17]. This finding underscores previous research asserting the protective role of physical activity against adverse health conditions linked to diabetes [12, 13, 10], broadening our understanding of the association between physical activity and major chronic health conditions in the context of diabetes.

However, it is crucial to acknowledge the limitations of our study, which primarily stem from the reliance on self-reported data. The self-reported nature of this data might introduce measurement errors and biases, impacting the accuracy of the statistical models employed in our study. Further, the cross-sectional nature of the present study restricts it to drawing out associations rather than establishing causal relationships.

Notwithstanding these limitations, the results contribute substantially to our understanding of diabetes management. In the realm of diabetes, where management often depends on costly interventions [7], these findings reiterate the significance of low-cost interventions like physical activity. Furthermore, bolstered by similar assertions from previous research, these findings hold potential to inform targeted interventions advocating physical activity for improving the health status of individuals with diabetes.

Future directions of research could encompass a wider range of chronic health conditions associated with diabetes. Additionally, exploring other potential confounding factors such as income, geographical location, etc. could deepen our understanding related to the role of physical activity in diabetes management. By extending research to broader contexts and various population groups, we can impart significant advancements in our understanding of the role of physical activity in diabetes and associated health conditions.

Methods

Data Source

The data for this study were obtained from the Behavioral Risk Factor Surveillance System (BRFSS), which is an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC). The dataset used in this study consisted of responses collected in the year 2015. The BRFSS survey collects information on health-related risk behaviors, chronic health conditions, and the use of preventative services from over 400,000 Americans each year. The dataset used in this study included 253,680 responses, with 22 features related to diabetes-related factors and chronic health conditions.

Data Preprocessing

The original dataset was provided in a CSV file format. Prior to analysis, the dataset was loaded into Python using the Pandas library. The data cleaning process involved the removal of any rows with missing values, resulting in a clean dataset of 253,680 responses with no missing values.

Data Analysis

In order to investigate the association between physical activity and chronic health conditions among individuals with diabetes, logistic regression models were utilized. Specifically, three logistic regression models were fitted to examine the relationship between physical activity and three chronic health conditions: high blood pressure, high cholesterol, and coronary heart disease. The models were built using the Statsmodels package in Python. Each model included physical activity as the main predictor variable, while controlling for potential confounding factors such as age, sex, BMI, and smoking status.

For each logistic regression model, the association between physical activity and the specific chronic health condition was assessed by examining the coefficients and p-values of the predictor variable. Additional statistical measures, such as the proportion of explained variance (pseudo R-squared), were also calculated to evaluate the accuracy of the models. The analysis results were saved in separate pickle files for further analysis and reporting.

It is important to note that the accuracy of the statistical models employed in this study was relatively low. This may be attributed to the use of self-reported data, which can introduce measurement errors and biases. However, despite these limitations, the findings from this analysis contribute to our understanding of the association between physical activity and chronic health conditions in individuals with diabetes.

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:

```python

import pandas as pd
import numpy as np

# Import Data
df = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.csv")

# Create an output file
with open('data_exploration.txt', 'w') as f:

    # Data Size
    f.write("# Data Size\n")
    f.write(f'Rows: {df.shape[0]}, Columns: {df.shape[1]}\n\n')

    # Summary Statistics
    f.write("# Summary Statistics\n")
    stats = df.describe(include='all')
    f.write(stats.to_string())
    f.write('\n\n')

    # Categorical Variables
    categorical_variables = df.select_dtypes(include=['int', 'bool'])
    f.write("# Categorical Variables\n")
    for var in categorical_variables:
        f.write(f"`{var}`\n")
        f.write(df[var].value_counts().to_string())
        f.write('\n\n')

    #Missing values
    f.write("# Missing Values\n")
    missing = df.isnull().sum()
    f.write(missing.to_string())

f.close()

```

B.2 Code Description

The code performs data exploration on the given dataset.

First, the code reads the dataset from the file "diabetes_binary_health_indicators_BRFSS2015.csv" and stores it in a Pandas DataFrame, named df.

Next, the code generates an output file named "data_exploration.txt" and opens it for writing.

The code then proceeds to perform the following analysis steps:

1. Data Size: The code calculates and writes the number of rows and columns in the dataset using the shape attribute of the DataFrame.

2. Summary Statistics: The code calculates and writes the summary statistics of the dataset using the describe() function. This includes descriptive statistics like count, mean, standard deviation, minimum, maximum, and quartiles for each column in the dataset.

3. Categorical Variables: The code identifies the categorical variables in the dataset based on their data type (integer or boolean) using the `select_dtypes()` function. For each categorical variable, the code writes the value counts, which represent the frequency of each category in the variable.

4. Missing Values: The code calculates and writes the number of missing values in each column of the dataset using the `isnull().sum()` function.

The output file ”data_exploration.txt” will contain the following information:

- Data Size: Number of rows and columns in the dataset. - Summary Statistics: Descriptive statistics for each column in the dataset. - Categorical Variables: Frequency of categories for each categorical variable. - Missing Values: Number of missing values in each column of the dataset.

This comprehensive data exploration provides an overview of the dataset's structure, summary statistics, categorical variable distributions, and missing data, which enables researchers to better understand the dataset and make informed decisions in subsequent data analysis processes.

B.3 Code Output

data_exploration.txt

```
# Data Size
Rows: 253680, Columns: 22

# Summary Statistics
       Diabetes_binary  HighBP  HighChol  CholCheck    BMI  Smoker  Stroke
    HeartDiseaseorAttack  PhysActivity  Fruits  Veggies  HvyAlcoholConsump
    AnyHealthcare  NoDocbcCost  GenHlth  MentHlth  PhysHlth  DiffWalk    Sex    Age
    Education  Income
count           253680  253680    253680     253680 253680  253680  253680
    253680        253680  253680   253680             253680         253680
    253680   253680    253680    253680    253680 253680 253680     253680  253680
mean            0.1393   0.429    0.4241     0.9627  28.38  0.4432 0.04057
    0.09419        0.7565 0.6343    0.8114             0.0562         0.9511
    0.08418    2.511     3.185     4.242    0.1682 0.4403  8.032       5.05   6.054
std             0.3463  0.4949    0.4942     0.1896  6.609  0.4968  0.1973
    0.2921        0.4292  0.4816   0.3912             0.2303         0.2158
    0.2777    1.068     7.413     8.718    0.3741 0.4964  3.054     0.9858   2.071
min                  0       0         0          0     12       0       0
```

```
    0             0       0        0                  0              0            0
    1         0         0         0      0      1          1       1
25%                  0       0         0          1     24       0       0
    0             1       1        1                  0              1            0
    2         0         0         0      0      6          4       5
50%                  0       0         0          1     27       0       0
    0             1       1        1                  0              1            0
    2         0         0         0      0      8          5       7
75%                  0       1         1          1     31       1       0
    0             1       1        1                  0              1            0
    3         2         3         0      1     10          6       8
max                  1       1         1          1     98       1       1
    1             1       1        1                  1              1            1
    5        30        30         1      1     13          6       8

# Categorical Variables
`Diabetes_binary`
Diabetes_binary
0    218334
1     35346

`HighBP`
HighBP
0    144851
1    108829

`HighChol`
HighChol
0    146089
1    107591

`CholCheck`
CholCheck
1    244210
0      9470

`BMI`
BMI
27    24606
26    20562
```

24	19550
25	17146
28	16545
23	15610
29	14890
30	14573
22	13643
31	12275
32	10474
21	9855
33	8948
34	7181
20	6327
35	5575
36	4633
37	4147
19	3968
38	3397
39	2911
40	2258
18	1803
41	1659
42	1639
43	1500
44	1043
45	819
17	776
46	750
47	622
48	484
49	416
50	372
16	348
51	253
53	237
52	215
55	169
15	132
54	113
56	109

57	86
58	71
79	66
60	63
87	61
77	55
59	54
75	52
71	49
81	49
73	47
84	44
62	43
14	41
82	37
61	35
63	34
92	32
89	28
64	24
13	21
65	19
74	16
67	15
70	15
72	14
68	14
66	13
95	12
69	9
98	7
12	6
76	3
88	2
83	2
80	2
96	1
85	1
91	1
86	1

```
90        1
78        1

`Smoker`
Smoker
0    141257
1    112423

`Stroke`
Stroke
0    243388
1     10292

`HeartDiseaseorAttack`
HeartDiseaseorAttack
0    229787
1     23893

`PhysActivity`
PhysActivity
1    191920
0     61760

`Fruits`
Fruits
1    160898
0     92782

`Veggies`
Veggies
1    205841
0     47839

`HvyAlcoholConsump`
HvyAlcoholConsump
0    239424
1     14256

`AnyHealthcare`
AnyHealthcare
```

```
1    241263
0     12417

`NoDocbcCost`
NoDocbcCost
0    232326
1     21354

`GenHlth`
GenHlth
2    89084
3    75646
1    45299
4    31570
5    12081

`MentHlth`
MentHlth
0     175680
2      13054
30     12088
5       9030
1       8538
3       7381
10      6373
15      5505
4       3789
20      3364
7       3100
25      1188
14      1167
6        988
8        639
12       398
28       327
21       227
29       158
18        97
9         91
16        88
```

```
27        79
22        63
17        54
26        45
11        41
13        41
23        38
24        33
19        16

`PhysHlth`
PhysHlth
0     160052
30     19400
2      14764
1      11388
3       8495
5       7622
10      5595
15      4916
4       4542
7       4538
20      3273
14      2587
25      1336
6       1330
8        809
21       663
12       578
28       522
29       215
9        179
18       152
16       112
27        99
17        96
24        72
22        70
26        69
13        68
```

```
11        60
23        56
19        22

`DiffWalk`
DiffWalk
0    211005
1     42675

`Sex`
Sex
0    141974
1    111706

`Age`
Age
9     33244
10    32194
8     30832
7     26314
11    23533
6     19819
13    17363
5     16157
12    15980
4     13823
3     11123
2      7598
1      5700

`Education`
Education
6    107325
5     69910
4     62750
3      9478
2      4043
1       174

`Income`
```

```
Income
8    90385
7    43219
6    36470
5    25883
4    20135
3    15994
2    11783
1     9811

# Missing Values
Diabetes_binary         0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
```

C Data Analysis

C.1 Code

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

```

# IMPORT
import pandas as pd
from statsmodels.formula.api import logit
import pickle

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

# DATASET PREPARATIONS
# No dataset preparations are needed.

# DESCRIPTIVE STATISTICS
# Creating Table 0: Descriptive statistics of Physical
↪   Activity and Chronic Health Conditions stratified by
↪   Whether Having Diabetes
desc_stats =
↪   df.groupby('Diabetes_binary').agg({"PhysActivity": "mean",
                                                "HighBP":
                                                ↪   "mean",
                                                "HighChol":
                                                ↪   "mean",
                                                
                                                ↪   "HeartDiseaseorAttack":
                                                ↪   "mean"}).reset_index()
desc_stats['Diabetes_binary'] =
↪   desc_stats['Diabetes_binary'].replace([0,1],['No
↪   Diabetes', 'Diabetes'])
desc_stats.set_index('Diabetes_binary', inplace=True)
desc_stats.to_pickle('table_0.pkl')

# PREPROCESSING
# No preprocessing is needed, because the data is already in a
↪   structured format with binary variables and continuous
↪   variables are not being used in the analysis.

# ANALYSIS

```

```python
## Table 1: "Association between physical activity and high blood pressure in individuals with diabetes"
model_highBP = logit("HighBP ~ PhysActivity + Age + Sex + BMI + Smoker", data=df[df["Diabetes_binary"]==1]).fit()
summary_df = model_highBP.summary2().tables[1]
df1 = summary_df.loc[['PhysActivity'], :]
df1.to_pickle('table_1.pkl')

## Table 2: "Association between physical activity and high cholesterol in individuals with diabetes"
model_highChol = logit("HighChol ~ PhysActivity + Age + Sex + BMI + Smoker", data=df[df["Diabetes_binary"]==1]).fit()
summary_df = model_highChol.summary2().tables[1]
df2 = summary_df.loc[['PhysActivity'], :]
df2.to_pickle('table_2.pkl')

## Table 3: "Association between physical activity and coronary heart disease in individuals with diabetes"
model_heartDisease = logit("HeartDiseaseorAttack ~ PhysActivity + Age + Sex + BMI + Smoker", data=df[df["Diabetes_binary"]==1]).fit()
summary_df = model_heartDisease.summary2().tables[1]
df3 = summary_df.loc[['PhysActivity'], :]
df3.to_pickle('table_3.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
 'Total number of observations': len(df),
 'Accuracy of high blood pressure model': model_highBP.prsquared,
 'Accuracy of high cholesterol model': model_highChol.prsquared,
 'Accuracy of coronary heart disease model': model_heartDisease.prsquared
}

with open('additional_results.pkl', 'wb') as f:
 pickle.dump(additional_results, f)

```

C.2 Code Description

The provided code performs an analysis on a dataset containing diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The analysis focuses on examining the association between physical activity and several chronic health conditions (high blood pressure, high cholesterol, coronary heart disease) in individuals with diabetes.

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

Next, descriptive statistics are computed for the variables "PhysActivity" (physical activity), "HighBP" (high blood pressure), "HighChol" (high cholesterol), and "HeartDiseaseorAttack" (coronary heart disease) stratified by whether an individual has diabetes or not. The descriptive statistics are saved as Table 0 in a pickle file.

The analysis is then performed using logistic regression models. Three separate models are fitted to examine the association between physical activity and each of the three chronic health conditions (high blood pressure, high cholesterol, coronary heart disease), considering only individuals with diabetes.

For each model, the code computes the logistic regression model using the "logit" function from the statsmodels library. The independent variables include "PhysActivity" (physical activity), "Age", "Sex", "BMI", and "Smoker". The dependent variable is "HighBP" for the high blood pressure model, "HighChol" for the high cholesterol model, and "HeartDiseaseorAttack" for the coronary heart disease model.

The code saves the results of each model, specifically the coefficient estimates, standard errors, p-values, and confidence intervals, for the "PhysActivity" variable as Table 1, Table 2, and Table 3, respectively, in separate pickle files.

Additionally, the code computes and saves additional results in the "additional_results.pkl" file. These results include the total number of observations in the dataset and the accuracy (pseudo R-squared) of each of the three logistic regression models.

In summary, the provided code performs an analysis to investigate the association between physical activity and three chronic health conditions (high blood pressure, high cholesterol, coronary heart disease) in individuals with diabetes. Logistic regression models are used to estimate these associations, and the results are saved in separate tables and additional results files.

C.3 Code Output

table_0.pkl

```
                 PhysActivity  HighBP  HighChol  HeartDiseaseorAttack
Diabetes_binary
No Diabetes            0.7769  0.3766    0.3843               0.07335
Diabetes               0.6305  0.7527    0.6701                0.2229
```

table_1.pkl

```
               Coef. Std.Err.      z      P>|z| [0.025  0.975]
PhysActivity -0.1718  0.02717 -6.322  2.587e-10 -0.225 -0.1185
```

table_2.pkl

```
               Coef. Std.Err.      z      P>|z|  [0.025   0.975]
PhysActivity -0.1175  0.02411 -4.873  1.102e-06 -0.1647 -0.07022
```

table_3.pkl

```
               Coef. Std.Err.      z      P>|z|  [0.025  0.975]
PhysActivity -0.3082  0.02718 -11.34  8.548e-30 -0.3615 -0.2549
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'Accuracy of high blood pressure model': 0.04641            ,
    'Accuracy of high cholesterol model': 0.006661            ,
    'Accuracy of coronary heart disease model': 0.05035            ,
}
```

D LaTeX Table Design

D.1 Code

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

```


# IMPORT
```

```python
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]]]

# 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': ('Phys. Act.', 'Physical Activity in past 30 days (0=no, 1=yes)'),
 'HighBP': ('High BP', 'High Blood Pressure (0=no, 1=yes)'),
 'HighChol': ('High Chol.', 'High Cholesterol (0=no, 1=yes)'),
 'HeartDiseaseorAttack': ('Heart Dis./Att.', 'Coronary heart disease (CHD) or myocardial infarction (MI), (0=no, 1=yes)'),
 'P>|z|':('P-value', 'P-value of the logistic regression model'),
 'z': ('z-score', 'Z-score for the coefficient in the logistic regression model'),
 'Coef.': ('Coeff.', 'Estimated model coefficient'),
 'Std.Err.': ('Std Err.', 'Standard error for the estimated coefficient'),
 '[0.025': ('CI Lower', '95% Confidence Interval Lower Bound'),
 '0.975]': ('CI Upper', '95% Confidence Interval Upper Bound')
}

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

# RENAME ROWS AND COLUMNS
```

```python
mapping = {k: v for k, v in shared_mapping.items() if k in df.columns or k in df.index}
abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(columns=abbrs_to_names, index=abbrs_to_names)

# Save as latex:
to_latex_with_note(df, 'table_0.tex',
                   caption="Descriptive Statistics of Physical Activity and Chronic Health Conditions for both Diabetes and Non-Diabetes Individuals",
                   label='table:diabetes_comparison',
                   note="Values represent the proportions of individuals",
                   legend=legend)

# TABLE 1:
df = pd.read_pickle('table_1.pkl')

# FORMAT VALUES
df['P>|z|'] = df['P>|z|'].apply(format_p_value)

# RENAME COLUMN AND ROW NAMES
mapping = {k: v for k, v in shared_mapping.items() if k in df.columns or k in df.index}
abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(columns=abbrs_to_names, index=abbrs_to_names)

# Save as Latex
to_latex_with_note(df, 'table_1.tex',
                   caption="Association between Physical Activity and High BP in Individuals with Diabetes",
                                                                                    label='table:physical_activity_high_blood_pressure',
                   note="Values represent logistic regression coefficients. P-values are two-sided.",
                   legend=legend)


```

```python
# TABLE 2:
df = pd.read_pickle('table_2.pkl')

# FORMAT VALUES
df['P>|z|'] = df['P>|z|'].apply(format_p_value)

# RENAME COLUMN AND ROW NAMES
mapping = {k: v for k, v in shared_mapping.items() if k in df.columns or k in df.index}
abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(columns=abbrs_to_names, index=abbrs_to_names)

# Save as Latex
to_latex_with_note(df, 'table_2.tex',
                   caption="Association between Physical Activity and High Chol. in Individuals with Diabetes",
                   label='table:physical_activity_high_cholesterol',
                   note="Values represent logistic regression coefficients. P-values are two-sided.",
                   legend=legend)


# TABLE 3:
df = pd.read_pickle('table_3.pkl')

# FORMAT VALUES
df['P>|z|'] = df['P>|z|'].apply(format_p_value)

# RENAME COLUMN AND ROW NAMES
mapping = {k: v for k, v in shared_mapping.items() if k in df.columns or k in df.index}
abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(columns=abbrs_to_names, index=abbrs_to_names)


# Save as Latex
to_latex_with_note(df, 'table_3.tex',
                   caption="Association between Physical
                   ↪  Activity and Heart Dis./Att. in
                   ↪  Individuals with Diabetes",

                   ↪  label='table:physical_activity_heart_disease',
                   note="Values represent logistic regression
                   ↪  coefficients. P-values are two-sided.",
                   legend=legend)

```

D.2 Code Output

table_0.tex

```
\begin{table}[h]
\caption{Descriptive Statistics of Physical Activity and Chronic Health
    Conditions for both Diabetes and Non-Diabetes Individuals}
\label{table:diabetes_comparison}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrrrr}
\toprule
 & Phys. Act. & High BP & High Chol. & Heart Dis./Att. \\
Diabetes\_binary &  &  &  &  \\
\midrule
\textbf{No Diabetes} & 0.777 & 0.377 & 0.384 & 0.0734 \\
\textbf{Diabetes} & 0.631 & 0.753 & 0.67 & 0.223 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Values represent the proportions of individuals
\item \textbf{Phys. Act.}: Physical Activity in past 30 days (0=no, 1=yes)
\item \textbf{High BP}: High Blood Pressure (0=no, 1=yes)
\item \textbf{High Chol.}: High Cholesterol (0=no, 1=yes)
\item \textbf{Heart Dis./Att.}: Coronary heart disease (CHD) or myocardial
    infarction (MI), (0=no, 1=yes)
\end{tablenotes}
\end{threeparttable}
```

```
\end{table}
```

table_1.tex

```
\begin{table}[h]
\caption{Association between Physical Activity and High BP in Individuals with
    Diabetes}
\label{table:physical_activity_high_blood_pressure}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coeff. & Std Err. & z-score & P-value & CI Lower & CI Upper \\
\midrule
\textbf{Phys. Act.} & -0.172 & 0.0272 & -6.32 & $<$1e-06 & -0.225 & -0.119 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Values represent logistic regression coefficients. P-values are two-sided.
\item \textbf{Phys. Act.}: Physical Activity in past 30 days (0=no, 1=yes)
\item \textbf{P-value}: P-value of the logistic regression model
\item \textbf{z-score}: Z-score for the coefficient in the logistic regression
    model
\item \textbf{Coeff.}: Estimated model coefficient
\item \textbf{Std Err.}: Standard error for the estimated coefficient
\item \textbf{CI Lower}: 95\% Confidence Interval Lower Bound
\item \textbf{CI Upper}: 95\% Confidence Interval Upper Bound
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_2.tex

```
\begin{table}[h]
\caption{Association between Physical Activity and High Chol. in Individuals
    with Diabetes}
```

```
\label{table:physical_activity_high_cholesterol}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coeff. & Std Err. & z-score & P-value & CI Lower & CI Upper \\
\midrule
\textbf{Phys. Act.} & -0.117 & 0.0241 & -4.87 & 1.1e-06 & -0.165 & -0.0702 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Values represent logistic regression coefficients. P-values are two-sided.
\item \textbf{Phys. Act.}: Physical Activity in past 30 days (0=no, 1=yes)
\item \textbf{P-value}: P-value of the logistic regression model
\item \textbf{z-score}: Z-score for the coefficient in the logistic regression
    model
\item \textbf{Coeff.}: Estimated model coefficient
\item \textbf{Std Err.}: Standard error for the estimated coefficient
\item \textbf{CI Lower}: 95\% Confidence Interval Lower Bound
\item \textbf{CI Upper}: 95\% Confidence Interval Upper Bound
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_3.tex

```
\begin{table}[h]
\caption{Association between Physical Activity and Heart Dis./Att. in
    Individuals with Diabetes}
\label{table:physical_activity_heart_disease}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coeff. & Std Err. & z-score & P-value & CI Lower & CI Upper \\
\midrule
\textbf{Phys. Act.} & -0.308 & 0.0272 & -11.3 & $<$1e-06 & -0.361 & -0.255 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Values represent logistic regression coefficients. P-values are two-sided.
\item \textbf{Phys. Act.}: Physical Activity in past 30 days (0=no, 1=yes)
\item \textbf{P-value}: P-value of the logistic regression model
\item \textbf{z-score}: Z-score for the coefficient in the logistic regression
    model
\item \textbf{Coeff.}: Estimated model coefficient
\item \textbf{Std Err.}: Standard error for the estimated coefficient
\item \textbf{CI Lower}: 95\% Confidence Interval Lower Bound
\item \textbf{CI Upper}: 95\% Confidence Interval Upper Bound
\end{tablenotes}
\end{threeparttable}
\end{table}
```

References

[1] Gisela Cipullo Moreira, J. P. Cipullo, L. A. Ciorlia, C. Cesarino, and J. Vilela-Martin. Prevalence of metabolic syndrome: Association with risk factors and cardiovascular complications in an urban population. *PLoS ONE*, 9, 2014.

[2] Kylie Hill, Kylie Hill, Kylie Hill, P. Gardiner, V. Cavalheri, V. Cavalheri, S. Jenkins, S. Jenkins, S. Jenkins, G. Healy, and G. Healy. Physical activity and sedentary behaviour: applying lessons to chronic obstructive pulmonary disease. *Internal Medicine Journal*, 45, 2015.

[3] W. Franssen, G. H. Franssen, J. Spaas, F. Solmi, and B. O. Eijnde. Can consumer wearable activity tracker-based interventions improve physical activity and cardiometabolic health in patients with chronic diseases? a systematic review and meta-analysis of randomised controlled trials. *The International Journal of Behavioral Nutrition and Physical Activity*, 17, 2020.

[4] R. Rubin, T. Wadden, J. Bahnson, G. Blackburn, F. Brancati, G. Bray, Mace Coday, S. Crow, J. M. Curtis, G. Dutton, Caitlin M. Egan,

M. Evans, L. Ewing, L. Faulconbridge, J. Foreyt, S. Gaussoin, E. Gregg, H. Hazuda, James O Hill, E. Horton, V. Hubbard, J. Jakicic, R. Jeffery, K. Johnson, S. Kahn, W. Knowler, W. Lang, C. Lewis, M. Montez, Anne Murillo, D. Nathan, Jennifer Patricio, A. Peters, X. Pi-Sunyer, H. Pownall, W. Rejeski, Renate Rosenthal, Valerie Ruelas, Katie Toledo, B. van Dorsten, M. Vitolins, D. Williamson, R. Wing, S. Yanovski, and Ping Zhang. Impact of intensive lifestyle intervention on depression and health-related quality of life in type 2 diabetes: The look ahead trial. *Diabetes Care*, 37:1544 – 1553, 2014.

[5] Saee Hamine, E. Gerth-Guyette, D. Faulx, B. Green, and A. Ginsburg. Impact of mhealth chronic disease management on treatment adherence and patient outcomes: A systematic review. *Journal of Medical Internet Research*, 17, 2015.

[6] C. Rosenfeld. Sexdependent differences in voluntary physical activity. *Journal of Neuroscience Research*, 95, 2017.

[7] Iswarya Santhanakrishnan, S. Lakshminarayanan, and S. Kar. Factors affecting compliance to management of diabetes in urban health center of a tertiary care teaching hospital of south india. *Journal of Natural Science, Biology, and Medicine*, 5:365 – 368, 2014.

[8] W. Polonsky, D. Hessler, K. Ruedy, and R. Beck. The impact of continuous glucose monitoring on markers of quality of life in adults with type 1 diabetes: Further findings from the diamond randomized clinical trial. *Diabetes Care*, 40:736 – 741, 2017.

[9] D. Thom, A. Ghorob, D. Hessler, Diana De Vore, Ellen H. Chen, and Thomas Bodenheimer. Impact of peer health coaching on glycemic control in low-income patients with diabetes: A randomized controlled trial. *The Annals of Family Medicine*, 11:137 – 144, 2013.

[10] M. Hamer and E. Stamatakis. Low-dose physical activity attenuates cardiovascular disease mortality in men and women with clustered metabolic risk factors. *Circulation: Cardiovascular Quality and Outcomes*, 5:494499, 2012.

[11] Samson Y. Gebreab, S. Davis, J. Symanzik, G. Mensah, G. Gibbons, and A. Diez-Roux. Geographic variations in cardiovascular health in the united states: Contributions of state- and individual-level factors. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 4, 2015.

[12] R. Stein, C. Rockman, Yu Guo, M. Adelman, T. Riles, W. Hiatt, and J. Berger. Association between physical activity and peripheral artery disease and carotid artery stenosis in a self-referred population of 3 million adults. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 35:206212, 2015.

[13] Susan M. Devaraj, A. Kriska, T. Orchard, Rachel G. Miller, and T. Costacou. Cardiovascular health in early adulthood predicts the development of coronary heart disease in individuals with type 1 diabetes: 25 year follow-up from the pittsburgh epidemiology of diabetes complications study. *Diabetologia*, 64:571 – 580, 2020.

[14] H. Bowles, J. Morrow, Bruce Leonard, M. Hawkins, and P. Couzelis. The association between physical activity behavior and commonly reported barriers in a worksite population. *Research Quarterly for Exercise and Sport*, 73:464 – 470, 2002.

[15] A. Fretts, B. Howard, B. McKnight, G. Duncan, S. Beresford, M. Mete, Ying Zhang, and D. Siscovick. Lifes simple 7 and incidence of diabetes among american indians: The strong heart family study. *Diabetes Care*, 37:2240 – 2245, 2014.

[16] Jing Fang, Quanhe Yang, Yuling Hong, and Fleetwood Loustalot. Status of cardiovascular health among adult americans in the 50 states and the district of columbia, 2009. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 1, 2012.

[17] K. Heslin and Jeffrey E. Hall. Sexual orientation disparities in risk factors for adverse covid-19related outcomes, by race/ethnicity behavioral risk factor surveillance system, united states, 20172019. *Morbidity and Mortality Weekly Report*, 70:149 – 154, 2021.

[18] Minggen Lu and Wei Yang. Multivariate logistic regression analysis of complex survey data with application to brfss data. *Journal of Data Science*, 2012.

[19] Zelalem G. Dessie and T. Zewotir. Mortality-related risk factors of covid-19: a systematic review and meta-analysis of 42 studies and 423,117 patients. *BMC Infectious Diseases*, 21, 2021.

[20] J. Cawley, Aparna Soni, and K. Simon. Third year of survey data shows continuing benefits of medicaid expansions for low-income child-

less adults in the u.s. *Journal of General Internal Medicine*, 33:1495–1497, 2018.

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

[22] E. Maddaloni, N. Lessan, A. Al Tikriti, R. Buzzetti, P. Pozzilli, and M. Barakat. Latent autoimmune diabetes in adults in the united arab emirates: Clinical features and factors related to insulin-requirement. *PLoS ONE*, 10, 2015.

[23] Ronaldo Iachan, Carol Pierannunzi, Kristie Healey, K. Greenlund, and Machell Town. National weighting of data from the behavioral risk factor surveillance system (brfss). *BMC Medical Research Methodology*, 16, 2016.

[24] C. van Walraven, P. Austin, Alison Jennings, H. Quan, and A. Forster. A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical Care*, 47:626–633, 2009.

[25] Sarah Partridge, J. Balayla, C. Holcroft, and H. Abenhaim. Inadequate prenatal care utilization and risks of infant mortality and poor birth outcome: A retrospective analysis of 28,729,765 u.s. deliveries over 8 years. *American Journal of Perinatology*, 29:787 – 794, 2012.

[26] Jeffrey A. Johnson, S. Majumdar, S. Simpson, and E. Toth. Decreased mortality associated with the use of metformin compared with sulfonylurea monotherapy in type 2 diabetes. *Diabetes care*, 25 12:2244–8, 2002.

[27] L. Anderson, R. Pfeiffer, O. Landgren, S. Gadalla, Sonja I. Berndt, and E. Engels. Risks of myeloid malignancies in patients with autoimmune conditions. *British Journal of Cancer*, 100:822 – 828, 2009.

[28] R. Plotnikoff, L. Taylor, P. Wilson, K. Courneya, R. Sigal, N. Birkett, K. Raine, and L. Svenson. Factors associated with physical activity in canadian adults with diabetes. *Medicine and science in sports and exercise*, 38 8:1526–34, 2006.