Exercise as a Modulator of Chronic Disease Risk in Diabetes

data-to-paper

February 22, 2024

Abstract

Diabetes stands as a significant contributor to the global disease burden, often co-existing with other chronic conditions that compound morbidity and mortality. This study sought to address the research gap concerning the extent to which physical activity can mitigate the risks of chronic diseases such as high blood pressure, high cholesterol, and heart disease specifically within the diabetic population. Using a comprehensive 2015 CDC's Behavioral Risk Factor Surveillance System dataset encompassing over a quarter-million individuals, we employed statistical models to assess health outcomes in relation to exercise regimens. Results consistently highlighted a beneficial link between regular physical activity and reduced occurrence of aforementioned comorbid conditions among diabetic adults. Notably, physical activity appeared as a significant factor; however, its moderate explanatory power on disease risks underlines the multifactorial nature of chronic disease emergence and progression. While the study's cross-sectional framework limits a definitive causal inference, our findings significantly underscore the relevance of integrating physical activity into diabetes management strategies, alongside other metabolic and lifestyle interventions. These insights reinforce the intersection of public health and clinical practice in managing chronic disease risk through lifestyle modifications.

Introduction

The escalation of type-2 diabetes prevalence worldwide presents a formidable challenge to global health, individual quality of life, healthcare systems, and economic productivity [1]. Diabetes often coexists with other chronic conditions such as high blood pressure and heart diseases, augmenting its associated morbidity and mortality rates [2]. Moreover, it has far-reaching impacts like financial pressures on healthcare systems and productivity losses due to increased morbidity and premature mortality. As such, diabetes prevention and management strategies are imperative. In this regard, one overarching lifestyle modification strategy, physical activity, offers immense potential for managing diabetes and its associated health risks [3, 4, 5].

Though the importance of physical activity in managing diabetes is a well-researched area, a critical lacuna exists in understanding the extent to which it influences the risks associated with chronic diseases specifically among people diagnosed with type-2 diabetes. Several studies have already established the beneficial role of physical activity in blood glucose control and reducing the risk of type-2 diabetes onset [3, 4]. Simultaneously, it is suggested to suppress concentrations of risk factors like 17 alpha-estradiol in women, providing primary prevention of chronic diseases [5]. However, how this protection extends to affect the likelihood of co-existing conditions such as high blood pressure, high cholesterol, and heart diseases among those already living with diabetes remains less explored [6, 7].

Addressing this research gap is of paramount importance. Filling this gap affords valuable insights into managing diabetes and associated health risks more comprehensively. This, in turn, could assist in tailoring personalized interventions, forecasting disease outcomes more accurately, and directing future research efforts more effectively, ultimately benefiting those living with diabetes.

To this end, this study utilizes the comprehensive dataset from the 2015 CDC's Behavioral Risk Factor Surveillance System (BRFSS) as a robust platform for investigating this clinical question [8]. We chose a cross-sectional study design as it allows the study of the concurrent impact of physical activity on chronic disease risks among diabetic individuals. We further relied on logistic regression models to scrutinize health outcomes concerning physical activity regimens among diabetic adults [9]. We controlled for potential confounders like age, sex, body mass index, and smoking status, to ensure a precise estimation of the effect of physical activity [10]. This approach allowed us to discern whether regular physical activity can reduce the occurrence of high blood pressure, high cholesterol, and heart disease among adults diagnosed with diabetes, thereby contributing valuable insights to the body of knowledge in this field.

Results

First, to understand the overall prevalence of physical activity and certain chronic conditions among individuals with and without diabetes, we conducted a comparative analysis. Descriptive statistics, illustrated in Table 1, reveal that the mean frequency of physical activity is lower among respondents with diabetes (0.631) compared to those without diabetes (0.777). High blood pressure and high cholesterol were more commonly reported among individuals with diabetes (0.753 and 0.67, respectively) than among non-diabetic individuals (0.377 and 0.384). Moreover, the mean reported incidence of heart disease among persons with diabetes (0.223) exceeds that reported by individuals without the condition (0.0734).

Table 1: Descriptive statistics of Physical Activity and Chronic Health Conditions stratified by Diabetes status

Diabetes Status (0=No, 1=Yes)	Phys. Act.	High BP	High Chol.	Heart Disease
No	0.777	0.377	0.384	0.0734
Yes	0.631	0.753	0.67	0.223

Values represent frequency distributions
Phys. Act.: 0: Inactive, 1: Active
High BP: 0: No, 1: Yes
High Chol.: 0: No, 1: Yes
Heart Disease: 0: No, 1: Yes
No: Individuals without Diabetes

Then, to test the specific association between physical activity and high blood pressure in individuals diagnosed with diabetes, we performed a logistic regression analysis. As detailed in Table 2, engaging in physical activity was significantly associated with lower odds of reporting high blood pressure among the diabetic cohort. Physical activity was associated with a coefficient of -0.172 (P-value: $<10^{-6}$), suggesting lower likelihoods of high blood pressure for individuals who are active after accounting for age, sex, BMI, and smoking status. The 95% confidence interval for this association ranges from -0.225 to -0.119. However, the pseudo R-squared value of 0.04641 indicates that physical activity explains only a small proportion of the variation in high blood pressure among those with diabetes, indicating limited explanatory power.

Similarly, we explored the role of physical activity in influencing high cholesterol levels among individuals with diabetes. The findings, presented in Table 3, portray a negative association between physical activity and high cholesterol, with the coefficient estimated at -0.117 (P-value: $1.1\ 10^{-6}$). Nevertheless, the exceedingly low pseudo R-squared value of 0.006661 emphasizes that the model accounts for a very small fraction of the variance in

Table 2: Association between physical activity and high blood pressure among diabetics

	Coef.	Std.Err.	Z-score	P>|z|	[0.025	0.975]
Phys. Act.	-0.172	0.0272	-6.32	$<10^{-6}$	-0.225	-0.119

Values represent logistic regression coefficients
Phys. Act.: 0: Inactive, 1: Active
Z-score: Standard Score or Z-score is a metric that describes a values relationship to the mean of a group of values.

high cholesterol levels and underscores the caution needed when interpreting these results. The observed 95% confidence interval for this relationship extends from -0.165 to -0.0702.

Table 3: Association between physical activity and high cholesterol among diabetics

	Coef.	Std.Err.	Z-score	P>|z|	[0.025	0.975]
Phys. Act.	-0.117	0.0241	-4.87	$1.1\ 10^{-6}$	-0.165	-0.0702

Values represent logistic regression coefficients
Phys. Act.: 0: Inactive, 1: Active
Z-score: Standard Score or Z-score is a metric that describes a values relationship to the mean of a group of values.

Finally, to further evaluate the effect of physical activity on the occurrence of heart disease among those with diabetes, a logistic regression model was applied (Table 4). Being physically active was significantly connected with a decreased probability of developing heart disease, with a coefficient of -0.308 (P-value: $<10^{-6}$). The range of the 95% confidence interval for this association was -0.361 to -0.255. The model's pseudo R-squared value of 0.05035 points to its limited capacity in explaining the variance observed in coronary heart disease amongst diabetic patients, similar to that observed in the high blood pressure model.

In summary, these results demonstrate statistically significant negative associations between physical activity and the likelihood of chronic conditions such as high blood pressure, high cholesterol, and heart disease in diabetic adults, suggesting a potential protective role for physical activity. Despite the compelling nature of this evidence, the limited pseudo R-squared values and the cross-sectional design of the study necessitate a cautious interpretation, with the findings indicative of associations rather than causal rela-

Table 4: Association between physical activity and heart disease among diabetics

	Coef.	Std.Err.	Z-score	P>|z|	[0.025	0.975]
Phys. Act.	-0.308	0.0272	-11.3	$<10^{-6}$	-0.361	-0.255

Values represent logistic regression coefficients
Phys. Act.: 0: Inactive, 1: Active
Z-score: Standard Score or Z-score is a metric that describes a values relationship to the mean of a group of values.

tionships. Future prospective studies are required to elucidate the causality and underlying mechanisms of these observed associations.

Discussion

The growing prevalence of type-2 diabetes and its intertwining with other chronic diseases underscore the urgency for effective prevention and management strategies [1]. Among such strategies, physical activity has been suggested to hold significant potential for managing diabetes and associated health risks. However, this potential has been understudied specifically amongst the diabetic population [3, 4, 5].

In seeking to bridge this gap, this study utilized a robust dataset from the 2015 BRFSS survey and applied logistic regression models for assessing health outcomes of diabetic patients in relation to physical activity. We found that engaging in physical activity was significantly associated with decreased occurrence of high blood pressure, high cholesterol, and heart disease among adults with diabetes. However, the partial correlation statistics of our models suggest that the explanatory power of physical activity on these disease risks might be limited despite its significance. This is likely due to the multifactorial nature of chronic diseases, where a myriad of biological and environmental factors come into play [11].

Comparison of our findings with previous literature underscores their importance. Physical activity is broadly linked to multiple disease outcomes - a protective role that our study affirms in the context of diabetic adults [6, 7]. However, in contrast with some studies that have demonstrated a strong association between physical activity and disease outcomes, our results reveal a more moderate association [12]. This discrepancy could be attributed to differences in the population studied and the varied intensity and duration of physical activity involved.

Our results should be viewed in light of certain limitations. The primary restraint is the cross-sectional design, precluding any causal interpretations. Though we adjusted for potential confounders identified in the scientific literature, the possibility of residual confounding cannot be eliminated. Relying on self-reported measures also introduces potential bias that might have affected the results. Moreover, considering the binary or ordinal nature of the variables in the dataset, finer details regarding the intensity and duration of the physical activity could not be ascertained, which might have implications on interpreting the strength of the observed associations [5].

Our study opens avenues for future research to probe deeper into the cause-effect relationship between physical activity and chronic disease outcomes in diabetes. Longitudinal studies could offer the temporal framework required to better interpret the dynamicity of this relationship. Including more nuanced measures of physical activity and other potential confounders could enhance the rigor and explanatory power of future models.

In conclusion, our study underscores the protective role of physical activity in context to chronic disease risks among diabetic adults. Despite the limited explanatory capacity of physical activity, our findings hold substantial relevance for public health and clinical practice, affirming the inherent value of integrating physical activity into holistic diabetes management strategies. Further, they encourage comprehensive exploration of other influential factors contributing to this interplay, duly recognizing the intricacies of chronic disease progression.

Methods

Data Source

The study utilized a dataset compiled from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), conducted by the Centers for Disease Control and Prevention (CDC). This nationally representative survey collects information annually on health-related risk behaviors, chronic health conditions, and use of preventive services. Included in the dataset for this analysis were 253,680 individuals complete cases, with 22 features capturing demographic information, health behaviors, and health outcomes.

Data Preprocessing

The dataset utilized in this study required no further preprocessing efforts as all entries with missing data had been previously removed. Each feature in the dataset directly corresponds to survey responses or calculated variables derived from these responses. As our analyses focused on the associations of physical activity with specific chronic health conditions within individuals who have diabetes, we worked directly with the clean, structured dataset in binary and ordinal format as provided, without any necessity for additional preprocessing or data transformation.

Data Analysis

For the investigation of the influence of physical activity on chronic disease outcomes within the diabetic subsection of the dataset, we executed a series of statistical analyses. Specifically, we stratified the data based on diabetes status to examine descriptive statistics relevant to physical activity and chronic health conditions such as high blood pressure, high cholesterol, and coronary heart disease. Following the initial stratification, we implemented logistic regression models to assess the association between physical activity and each chronic condition, adjusting for potential confounding factors such as age, sex, body mass index, and smoking status. The selection of these covariates was informed by established risk factors for chronic diseases within the scientific literature. Through these logistic regression models, we derived odds ratios as measures of association, and we examined the models' pseudo R-squared values as indications of the goodness of fit to evaluate the explanatory power of the variables included. The practical significance and precision of our findings were inferred from the estimated associations and corresponding confidence intervals. All results were presented as computed from the models without additional manipulations.

Code Availability

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

References

[1] S. Nyberg, A. SinghManoux, A. SinghManoux, J. Pentti, J. Pentti, I. Madsen, S. Sabia, S. Sabia, Lars-Inge Alfredsson, L. Alfredsson, J. Bjorner, M. Borritz, H. Burr, M. Goldberg, M. Goldberg, K. Heikkil, M. Jokela, A. Knutsson, T. Lallukka, T. Lallukka, J. Lindbohm, M. Nielsen, M. Nordin, M. Nordin, T. Oksanen, J. Pejtersen, O. Rahkonen, R. Rugulies, M. Shipley, P. Sipil, S. Stenholm, S. Suominen,

S. Suominen, J. Vahtera, M. Virtanen, M. Virtanen, H. Westerlund, M. Zins, M. Zins, M. Hamer, G. Batty, G. Batty, M. Kivimki, and M. Kivimki. Association of healthy lifestyle with years lived without major chronic diseases. *JAMA Internal Medicine*, 180:1 – 10, 2020.

[2] S. Colberg, R. Sigal, J. Yardley, M. Riddell, D. Dunstan, P. Dempsey, E. Horton, Kristin Castorino, and D. Tate. Physical activity/exercise and diabetes: A position statement of the american diabetes association. *Diabetes Care*, 39:2065 – 2079, 2016.

[3] Dr.Kanchan Solanki. Exercise and type 2 diabetes. 2015.

[4] S. Colberg, R. Sigal, B. Fernhall, J. Regensteiner, B. Blissmer, R. Rubin, L. Chasan-Taber, A. Albright, and B. Braun. Exercise and type 2 diabetes. *Diabetes Care*, 33:2692 – 2696, 2010.

[5] J. Kruk. Physical activity in the prevention of the most frequent chronic diseases: an analysis of the recent evidence. *Asian Pacific journal of cancer prevention : APJCP*, 8 3:325–38, 2007.

[6] 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.

[7] S. Mora, N. Cook, J. Buring, P. Ridker, and I. Lee. Physical activity and reduced risk of cardiovascular events: Potential mediating mechanisms. *Circulation*, 116:2110–2118, 2007.

[8] 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.

[9] 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.

[10] Yonas Akalu, B. Ayelign, and M. Molla. Knowledge, attitude and practice towards covid-19 among chronic disease patients at addis zemen hospital, northwest ethiopia. *Infection and Drug Resistance*, 13:1949 – 1960, 2020.

[11] S. Barquera, Andrea Pedroza-Tobas, and C. Medina. Cardiovascular diseases in mega-countries: the challenges of the nutrition, physical activity and epidemiologic transitions, and the double burden of disease. *Current Opinion in Lipidology*, 27:329 – 344, 2016.

[12] 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.

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 to 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

# 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       0
```

```
            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
## 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.set_index('Diabetes_binary', inplace=True)
desc_stats.index.names = ['Diabetes Status (0=No, 1=Yes)']
desc_stats.index = desc_stats.index.map({0: 'No', 1: 'Yes'})  #
    ↪ Change the row index to categorical labels
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
## 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 code aims to analyze the relationship between physical activity and chronic health conditions in individuals with diabetes using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset.

After loading the dataset, the code calculates descriptive statistics of physical activity and chronic health conditions, stratified by whether individuals have diabetes or not. This information is saved in "table_0.pkl".

Next, three logistic regression models are fitted to assess the association between physical activity and three chronic health conditions (high blood pressure, high cholesterol, and coronary heart disease) in individuals with diabetes. Each model includes physical activity as the main predictor variable, along with other covariates such as age, sex, BMI, and smoking status. The results of these models are saved in three separate tables: "table_1.pkl", "table_2.pkl", and "table_3.pkl".

Additionally, the code calculates and saves some additional results. It calculates the total number of observations in the dataset and the percentage of variance explained (accuracy) by each model. These additional results are stored in a dictionary and saved in the "additional_results.pkl" file.

The code provides researchers with valuable information on the associations between physical activity and chronic health conditions in individuals with diabetes, contributing to the understanding of the importance of physical activity in managing and preventing these conditions.

C.3 Code Output

table_0.pkl

```
                    PhysActivity  HighBP  HighChol
   HeartDiseaseorAttack
```

```
Diabetes Status (0=No, 1=Yes)
No                                   0.7769  0.3766    0.3843
                 0.07335
Yes                                  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  2.59e-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  1.1e-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  8.55e-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
import pandas as pd
from my_utils import to_latex_with_note, is_str_in_df,
    ↪ split_mapping, AbbrToNameDef
from typing import Dict, Any, Tuple, Optional

# PREPARATION FOR ALL TABLES
shared_mapping: AbbrToNameDef = {
    'PhysActivity': ('Phys. Act.', '0: Inactive, 1: Active'),
```

```
    'HighBP': ('High BP', '0: No, 1: Yes'),
    'HighChol': ('High Chol.', '0: No, 1: Yes'),
    'HeartDiseaseorAttack': ('Heart Disease', '0: No, 1: Yes'),
    'No': ('No', 'Individuals without Diabetes'),
    'z': ('Z-score', 'Standard Score or Z-score is a metric
        ↪ that describes a values relationship to the mean of a
        ↪  group of values.')
}

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

# RENAME ROWS AND COLUMNS
mapping0 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df0, k))
abbrs_to_names0, legend0 = split_mapping(mapping0)
df0 = df0.rename(columns=abbrs_to_names0, index=abbrs_to_names0
    ↪ )

# SAVE AS LATEX:
to_latex_with_note(
    df0, 'table_0.tex',
    caption="Descriptive statistics of Physical Activity and
        ↪ Chronic Health Conditions stratified by Diabetes
        ↪ status",
    label='table:descriptive_statistics',
    note="Values represent frequency distributions",
    legend=legend0)

# TABLE 1:
df1 = pd.read_pickle('table_1.pkl')

# RENAME ROWS AND COLUMNS
mapping1 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df1, k))
abbrs_to_names1, legend1 = split_mapping(mapping1)
df1 = df1.rename(columns=abbrs_to_names1, index=abbrs_to_names1
    ↪ )

# SAVE AS LATEX:
to_latex_with_note(
    df1, 'table_1.tex',
    caption="Association between physical activity and high
        ↪ blood pressure among diabetics",
    label='table:association_highBP',
    note="Values represent logistic regression coefficients",
    legend=legend1)
```

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

# RENAME ROWS AND COLUMNS
mapping2 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df2, k))
abbrs_to_names2, legend2 = split_mapping(mapping2)
df2 = df2.rename(columns=abbrs_to_names2, index=abbrs_to_names2
    ↪ )

# SAVE AS LATEX:
to_latex_with_note(
    df2, 'table_2.tex',
    caption="Association between physical activity and high
        ↪ cholesterol among diabetics",
    label='table:association_highChol',
    note="Values represent logistic regression coefficients",
    legend=legend2)

# TABLE 3:
df3 = pd.read_pickle('table_3.pkl')

# RENAME ROWS AND COLUMNS
mapping3 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df3, k))
abbrs_to_names3, legend3 = split_mapping(mapping3)
df3 = df3.rename(columns=abbrs_to_names3, index=abbrs_to_names3
    ↪ )

# SAVE AS LATEX:
to_latex_with_note(
    df3, 'table_3.tex',
    caption="Association between physical activity and heart
        ↪ disease among diabetics",
    label='table:association_coronary',
    note="Values represent logistic regression coefficients",
    legend=legend3)
```

D.2 Provided Code

The code above is using the following provided functions:

```python
def to_latex_with_note(df, filename: str, caption: str, label:
    ↪ str, note: str = None, legend: Dict[str, str] = None, **
    ↪ kwargs):
    """
    Converts a DataFrame to a LaTeX table with optional note
        ↪ and legend added below the table.
```

```
    Parameters:
    - df, filename, caption, label: as in `df.to_latex`.
    - note (optional): Additional note below the table.
    - legend (optional): Dictionary mapping abbreviations to
        ↪ full names.
    - **kwargs: Additional arguments for `df.to_latex`.
    """

def is_str_in_df(df: pd.DataFrame, s: str):
    return any(s in level for level in getattr(df.index, '
        ↪ levels', [df.index]) + getattr(df.columns, 'levels',
        ↪ [df.columns]))

AbbrToNameDef = Dict[Any, Tuple[Optional[str], Optional[str]]]

def split_mapping(abbrs_to_names_and_definitions: AbbrToNameDef
    ↪ ):
    abbrs_to_names = {abbr: name for abbr, (name, definition)
        ↪ in abbrs_to_names_and_definitions.items() if name is
        ↪ not None}
    names_to_definitions = {name or abbr: definition for abbr,
        ↪ (name, definition) in abbrs_to_names_and_definitions.
        ↪ items() if definition is not None}
    return abbrs_to_names, names_to_definitions
```

D.3 Code Output

table_0.tex

```
% This latex table was generated from: `table_0.pkl`
\begin{table}[h]
\caption{Descriptive statistics of Physical Activity and
   Chronic Health Conditions stratified by Diabetes status}
\label{table:descriptive_statistics}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrrrr}
\toprule
 & Phys. Act. & High BP & High Chol. & Heart Disease \\
Diabetes Status (0=No, 1=Yes) &  &  &  &  \\
\midrule
\textbf{No} & 0.777 & 0.377 & 0.384 & 0.0734 \\
\textbf{Yes} & 0.631 & 0.753 & 0.67 & 0.223 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
```

```
\footnotesize
\item Values represent frequency distributions
\item \textbf{Phys. Act.}: 0: Inactive, 1: Active
\item \textbf{High BP}: 0: No, 1: Yes
\item \textbf{High Chol.}: 0: No, 1: Yes
\item \textbf{Heart Disease}: 0: No, 1: Yes
\item \textbf{No}: Individuals without Diabetes
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
% This latex table was generated from: `table_1.pkl`
\begin{table}[h]
\caption{Association between physical activity and high blood
   pressure among diabetics}
\label{table:association_highBP}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err. & Z-score & P$>$\textbar{}z\textbar{} &
    [0.025 & 0.975] \\
\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
\item \textbf{Phys. Act.}: 0: Inactive, 1: Active
\item \textbf{Z-score}: Standard Score or Z-score is a metric
   that describes a values relationship to the mean of a group
    of values.
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_2.tex

```
% This latex table was generated from: `table_2.pkl`
\begin{table}[h]
\caption{Association between physical activity and high
   cholesterol among diabetics}
\label{table:association_highChol}
\begin{threeparttable}
```

```
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err. & Z-score & P$>$\textbar{}z\textbar{} &
     [0.025 & 0.975] \\
\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
\item \textbf{Phys. Act.}: 0: Inactive, 1: Active
\item \textbf{Z-score}: Standard Score or Z-score is a metric
    that describes a values relationship to the mean of a group
     of values.
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_3.tex

```
% This latex table was generated from: `table_3.pkl`
\begin{table}[h]
\caption{Association between physical activity and heart
    disease among diabetics}
\label{table:association_coronary}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err. & Z-score & P$>$\textbar{}z\textbar{} &
     [0.025 & 0.975] \\
\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
\item \textbf{Phys. Act.}: 0: Inactive, 1: Active
\item \textbf{Z-score}: Standard Score or Z-score is a metric
    that describes a values relationship to the mean of a group
     of values.
\end{tablenotes}
```

```
\end{threeparttable}
\end{table}
```