Influence of Physical Activity and BMI on Diabetes Risk in an American Adult Population

data-to-paper

February 22, 2024

Abstract

Diabetes is an escalating public health issue, with lifestyle factors being crucial determinants of its risk and management. Despite mounting evidence on the influence of body weight and physical activity on diabetes, comprehensive analyses of their interplay are scarce. This study fills a research gap by evaluating how physical activity and body mass index (BMI) jointly contribute to diabetes prevalence. Using a large-scale dataset from the 2015 Behavioral Risk Factor Surveillance System, this study applied descriptive statistics and multiple linear regression to assess the lifestyle determinants of diabetes among more than 253,000 American adults. Results demonstrated that physical activity is inversely associated with diabetes likelihood, while BMI shows a direct correlation with increased risk, even when accounting for demographic disparities. Notably, gender differences manifested in diabetes rates with a somewhat elevated prevalence among males. These insights are critical, given the sparse data-driven analyses linking lifestyle factors to chronic disease risk across diverse populations. However, the cross-sectional study design does limit the ability to infer causality. Collectively, the results highlight lifestyle modification, including physical activity encouragement and weight management, as promising strategies for diabetes risk reduction, informing public health policies and individual preventive measures.

Introduction

Diabetes is a growing public health concern with over 34.2 million individuals in the United States alone affected by the disease [1]. With the disease playing a significant role in the onset of other serious health conditions such as heart disease and stroke, a comprehensive understanding of the diverse factors that influence diabetes risk at a population level becomes crucial [2]. While research has extensively examined the individual impacts of lifestyle factors such as Body Mass Index (BMI) and physical activity levels on diabetes prevalence [3, 4], a noteworthy research gap persists in exploring their combined contribution and interaction with demographic disparities in a large-scale population context [5, 6, 7].

Prevailing studies have tended to focus on the association between diabetes and either lifestyle or demographic factors, having paid less attention to the interaction of these determinants [8, 9]. The conclusion that heightened physical activity can have meaningful implications for glycemic control has been drawn [2, 4]. Similarly, studies have shown a correlation between increased BMI and a higher risk of developing the disease [3, 10]. Yet, the potentially nuanced interplay between these lifestyle factors and how they collectively relate to diabetes risk across different demographics remains an area warranting further exploration.

Addressing these unanswered research questions, the present study employs an extensive dataset from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) [11]. The BRFSS provides widespread health-related data annually, involving over 400,000 American adults, offering an extensive record of self-reported physical activity and BMI readings across diverse demographic groups [12, 13]. This unique dataset, distinct from those examined in previous analyses, allows for a more detailed investigation into the interplay of lifestyle factors, gender, and diabetes.

The robust analytical measures employed in our study include conducting a multiple linear regression model to evaluate the interaction between physical activity, BMI, and diabetes prevalence, along with demographic variables such as gender as controls [14, 15]. This in-depth approach allows us to delve deeper into the modulating impacts of lifestyle factors on diabetes risk across genders and highlights areas for potential public health intervention. Initial findings hint at significant associations between physical activity and BMI with diabetes prevalence, implying the potential health benefits of lifestyle modifications. Although the results corroborate certain findings from previous research, the simultaneous examination of multiple lifestyle and demographic factors offers a unique contribution to the field, providing novel insights into mitigating diabetes risk within the American population.

Results

First, to understand the baseline distribution of diabetes across different genders and its relationship with physical activity and Body Mass Index (BMI), we conducted an exploratory analysis. The descriptive statistics highlighted in Table 1 show distinct patterns between males and females. Females had a mean diabetes prevalence of 13%, compared to 15.2% in males. When considering physical activity, a higher proportion of males (77.2%) reported exercising in the past 30 days compared to females (74.4%). Furthermore, the mean BMI indicates slight differences, with females averaging at 28.1 and males higher at 28.7.

Table 1: Descriptive Statistics of Main Binary Variables and Body Mass Index Stratified by Gender

Sex	Female	Male
Mean Diabetes	0.13	0.152
Std. Diabetes	0.336	0.359
Mean Phys. Activity	0.744	0.772
Std. Phys. Activity	0.436	0.419
Mean BMI	28.1	28.7
Std. BMI	7.09	5.93

Descriptive statistics include mean and standard deviation of diabetes measure, physical activity, and body mass index.

Then, to test the association between physical activity levels and diabetes, adjusting for gender and BMI, we performed a multiple linear regression analysis. The model, as documented in Table 2, reveals the negative association between physical activity and the likelihood of diabetes. Specifically, each unit increase in physical activity is associated with a 0.0719 decrease in the probability of diabetes. Conversely, BMI exhibits a positive correlation with diabetes risk, indicating that each unit increase in BMI is linked to a 0.0106 increase in the likelihood of diabetes. The gender variable for males is associated with an increased diabetes risk, with a coefficient of 0.0179, suggesting that males have a slightly higher risk of diabetes compared to females after controlling for physical activity and BMI. The explanatory power of the model is modest, with an R-squared of 5.527%, suggesting that the variables included in the model can account for a small proportion of the variability in diabetes prevalence.

Finally, to ensure that the observed associations are not confounded by

Table 2: Multiple Linear Regression for Testing Association between Physical Activity Level and Diabetes, Adjusted by Age, Gender, and Body Mass Index

	Coefficient	P-value	CI (Lower)	CI (Higher)
Intercept	-0.115	$<10^{-6}$	-0.122	-0.109
Gender Male	0.0179	$<10^{-6}$	0.0152	0.0205
Phys. Activity	-0.0719	$<10^{-6}$	-0.075	-0.0688
Body Mass Index	0.0106	$<10^{-6}$	0.0104	0.0108

Table includes the results of a multiple linear regression model on the relationship between diabetes and physical activity, adjusted by gender, age, and body mass index.
Body Mass Index: Measure of body fat based on height and weight that applies to adult men and women
Phys. Activity: Physical activity in the past 30 days
Gender Male: Categorical variable, Male

demographic factors, the model includes sex as a control. With a total number of observations at 253,680, the findings offer robust statistics providing reliable insights into the relationships between diabetes prevalence, physical activity, and BMI.

In summary, these results from descriptive analyses and multiple linear regression suggest that physical activity is inversely associated, and BMI is directly associated with diabetes prevalence in an American adult population, after adjusting for gender. The large sample size lends credibility to the robustness of the results, despite the modest explanatory power of the model in terms of variability accounted for.

Discussion

The aim of this study was intricate, seeking to determine the connection between lifestyle factors - physical activity and Body Mass Index (BMI) - and their association with the prevalence of diabetes in context of demographic factors such as gender [16, 7, 5]. Previous studies have shown a clear correlation between physical activity and BMI with the risk of developing diabetes [6, 2]. However, this study diverges from previous research in its exploration of how these risk factors interplay with demographic characteristics to shape diabetes risk [5, 6, 7].

The methodology employed in this study involved extracting and analyzing data from the 2015 BRFSS dataset, employing descriptive statistics and performing multiple linear regression analysis [14, 15]. The outcomes reconfirm the inversely proportional relationship between physical activity and diabetes, and conversely, the direct correlation between BMI and the disease [5, 2]. The granular insight into demographic-specific influence on these lifestyle factors is a unique contribution of this study [8, 9].

The limitations of this study should also be considered while interpreting these results. The source data for the study was self-reported, which may introduce recall bias or misreporting, leading to potential incongruities in physical activity and BMI metrics. The cross-sectional design of the survey could limit the extent of causal interpretations. While the regression model accounted for some confounders, there may be other unconsidered confounding factors that influence diabetes prevalence, such as dietary habits and genetic predispositions. The somewhat modest explanatory power of the multiple regression model warrants caution, with the findings accounting for a relatively small proportion of the variability in diabetes prevalence.

The study's findings, coupled with contributions from other researchers, attest to the complexity of diabetes, and its regulation by factors such as physical activity, BMI, and gender [7, 16, 5]. While our model presents a modest explanatory power, it highlights the need for continued research to better capture the disease's multifactorial nature.

In conclusion, the joint examination of lifestyle and demographic factors as performed in this study emphasizes the pivotal role of constructive lifestyle modifications as part of a broader, personalized approach to mitigating diabetes risk at the population level. Notwithstanding the limitations of our study, these findings underscore the value of context-dependent approaches towards public health policy planning and disease management. Future research should consider implementing longitudinal study designs and diversifying the analytical scope to incorporate a comprehensive list of potentially influential factors. This could promote nuanced insights into diabetes risk and bolster strategies for its management and prevention.

Methods

Data Source

The data employed in this study was obtained from the Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. Managed by the Centers for Disease Control and Prevention (CDC), the BRFSS is an annual collection of health-related data gathered through telephone interviews with over 400,000 American adults. This dataset comprised responses on risk behaviors, chronic health conditions, and utilization of preventive services, formulated either as direct queries to the participants or as variables calculated based on responses.

Data Preprocessing

Our analysis utilized a dataset that had already undergone preprocessing to eliminate instances with missing values. The subsequent analytical process involved the transformation of categorical variables into a binary numerical format suitable for linear regression analysis. Specifically, dummy variables were created for certain demographic characteristics, including sex, age, educational attainment, and income level. These steps were necessary to allow for the inclusion of these variables in the regression model to control for potential confounding effects.

Data Analysis

The data analysis was orchestrated in two major steps. First, descriptive statistics were calculated for core variables including diabetes status, physical activity levels, and BMI, with results stratified by sex. These statistics provided an overview and summary of the base characteristics of the dataset, including means and standard deviations for the variables of interest.

Next, a multiple linear regression model was constructed to investigate the relationship between physical activity, BMI, and the likelihood of having diabetes. The model adjusted for sex as a confounding demographic factor. The beta coefficients, corresponding p-values, and confidence intervals were computed for each predictor within the regression, quantifying the association with the dependent variable, which was the binary diabetes indicator. The integrity of the model was assessed by the coefficient of determination, reflecting the proportion of variance for the dependent variable that is explained by the independent variables in the model. This step was crucial in elucidating the potential influence of physical activity and BMI on diabetes risk, whilst taking into account demographic variability.

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

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

[3] Ying Chen, Xiao-Ping Zhang, Jie Yuan, Bo zhi Cai, Xiao-Li Wang, Xiao li Wu, Yueqi Zhang, Xiao-Yi Zhang, T. Yin, Xiaohui Zhu, Y. Gu, S. Cui, Zhidong Lu, and Xiao ying Li. Association of body mass index and age with incident diabetes in chinese adults: a population-based cohort study. *BMJ Open*, 8, 2018.

[4] J. Kanaley, S. Colberg, Matthew H. Corcoran, S. Malin, N. Rodriguez, C. Crespo, J. Kirwan, and J. Zierath. Exercise/physical activity in individuals with type 2 diabetes: A consensus statement from the american college of sports medicine. *Medicine & Science in Sports & Exercise*, 54:353 – 368, 2022.

[5] K. V. Hjerkind, J. S. Stenehjem, and T. Nilsen. Adiposity, physical activity and risk of diabetes mellitus: prospective data from the population-based hunt study, norway. *BMJ Open*, 7, 2017.

[6] P. Boffetta, D. McLerran, Yu Chen, M. Inoue, R. Sinha, Jiang He, P. Gupta, S. Tsugane, F. Irie, A. Tamakoshi, Yu-Tang Gao, X. Shu, Renwei Wang, I. Tsuji, S. Kuriyama, K. Matsuo, H. Satoh, Chien-Jen Chen, Jian-Min Yuan, K. Yoo, H. Ahsan, W. Pan, D. Gu, M. Pednekar, S. Sasazuki, T. Sairenchi, Gong Yang, Y. Xiang, M. Nagai, Hideo Tanaka, Y. Nishino, S. You, W. Koh, Sue-Kyung Park, Chen-Yang Shen, M. Thornquist, D. Kang, Betsy Rolland, Ziding Feng, W. Zheng, and J. Potter. Body mass index and diabetes in asia: A cross-sectional pooled analysis of 900,000 individuals in the asia cohort consortium. *PLoS ONE*, 6, 2011.

[7] J. Seiglie, M. Marcus, Cara Ebert, Nikolaos Prodromidis, P. Geldsetzer, M. Theilmann, K. Agoudavi, Glennis Andall-Brereton, K. Aryal, B. Bicaba, P. Bovet, G. Brian, M. Dorobanu, G. Gathecha, M. Gurung,

D. Guwatudde, Mohamed Msaidi, C. Houehanou, D. Houinato, J. Jrgensen, G. Kagaruki, K. Karki, D. Labadarios, J. Martins, Mary T Mayige, R. Wong-McClure, J. K. Mwangi, Omar Mwalim, Bolormaa Norov, Sarah Quesnel-Crooks, Bahendeka K Silver, L. Sturua, Lindiwe Tsabedze, C. Wesseh, A. Stokes, R. Atun, J. Davies, S. Vollmer, T. Brnighausen, L. Jaacks, J. Meigs, D. Wexler, and J. Manne-Goehler. Diabetes prevalence and its relationship with education, wealth, and bmi in 29 low- and middle-income countries. *Diabetes Care*, 43:767 – 775, 2020.

[8] T. Pham, J. Carpenter, T. Morris, Manuj Sharma, and I. Petersen. Ethnic differences in the prevalence of type 2 diabetes diagnoses in the uk: Cross-sectional analysis of the health improvement network primary care database. *Clinical Epidemiology*, 11:1081 – 1088, 2019.

[9] Jaana Gustavsson, K. Mehlig, K. Leander, L. Lissner, L. Bjrck, A. Rosengren, and F. Nyberg. Fto genotype, physical activity, and coronary heart disease risk in swedish men and women. *Circulation: Cardiovascular Genetics*, 7:171177, 2014.

[10] J. Logue, Jeremy J. Walker, G. Leese, R. Lindsay, J. McKnight, A. Morris, S. Philip, S. Wild, and N. Sattar. Association between bmi measured within a year after diagnosis of type 2 diabetes and mortality. *Diabetes Care*, 36:887 – 893, 2013.

[11] 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.

[12] L. Aiello, R. Schifanella, D. Quercia, and Lucia Del Prete. Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Science*, 8, 2019.

[13] W. Davis, V. Parsons, D. Xie, N. Schenker, M. Town, T. Raghunathan, and E. Feuer. State-based estimates of mammography screening rates based on information from two health surveys. *Public Health Reports*, 125:567 – 578, 2010.

[14] D. Aune, T. Norat, M. Leitzmann, S. Tonstad, and L. Vatten. Physical activity and the risk of type 2 diabetes: a systematic review and doseresponse meta-analysis. *European Journal of Epidemiology*, 30:529–542, 2015.

[15] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112:1131 – 1146, 2015.

[16] F. Khanam, M. B. Hossain, S. Mistry, K. Afsana, and Mahfuzar Rahman. Prevalence and risk factors of cardiovascular diseases among bangladeshi adults: Findings from a cross-sectional study. *Journal of Epidemiology and Global Health*, 9:176 – 184, 2019.

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

# Load dataset
data = pd.read_csv('diabetes_binary_health_indicators_BRFSS2015
    ↪ .csv')

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

    ## Data size
    f.write('# Data Size\n')
    f.write('Number of Rows: {}\n'.format(data.shape[0]))
    f.write('Number of Columns: {}\n\n'.format(data.shape[1]))

    ## Summary Statistics for key numerical variables
    keys = ['BMI', 'GenHlth', 'MentHlth', 'PhysHlth', 'Age', '
        ↪ Education', 'Income']
    summary = data[keys].describe()

    f.write('# Summary Statistics\n')
    f.write('{}\n\n'.format(summary))

    ## Categorical Variables
    f.write('# Categorical Variables\n')
    categorical_columns = data.select_dtypes(include=['object',
        ↪ 'bool']).columns

    for column in categorical_columns:
        most_common = data[column].value_counts().idxmax()
        f.write('{}: Most common value is {}\n'.format(column,
            ↪ most_common))

    f.write('\n')

    ## Missing Values
    missing_values = data.isnull().sum()
    f.write('# Missing Values\n')
    f.write('{}\n\n'.format(missing_values))

    # Unknown/undefined values
    f.write('# Unknown/undefined values\n')
    for column in data.columns:
        count = data[data[column] == 99].shape[0]  # assuming
            ↪ 99 is the unknown/undefined value based on the
            ↪ dataset description
        f.write('{}: Number of unknown/undefined values: {}\n'.
            ↪ format(column, count))

f.close()
```

B.2 Code Description

The provided code performs a data exploration analysis on the diabetes related factors dataset. The analysis involves several steps to gain insights into the dataset's characteristics and identify any patterns or issues within the data.

First, the code loads the dataset into a Pandas DataFrame. Then, it opens an output file to store the results of the data exploration analysis.

The code starts by calculating the number of rows and columns in the dataset, which reflects the size of the data.

Next, summary statistics are computed for key numerical variables, including BMI (Body Mass Index), self-reported health (GenHlth), number of days with poor mental health (MentHlth), number of days with poor physical health (PhysHlth), age, education level, and income. These statistics provide an overview of the distribution of these variables in terms of measures such as mean, standard deviation, minimum, maximum, and quartiles.

The code then focuses on categorical variables in the dataset. It identifies the most common value for each categorical column, providing insights into the prevalence of different values within each category.

The analysis also checks for missing values in the dataset. The code calculates the count of missing values for each column and reports it in the output file. This information helps identify the completeness of the dataset and any potential issues related to missing data.

Finally, the code identifies unknown or undefined values in the dataset. It iterates over each column and counts the occurrences of a specific value (assumed to be 99 in this case) that indicates an unknown or undefined value. This information can be crucial in understanding the quality and reliability of the data.

The code writes all the analysis results into the "data_exploration.txt" file. The file includes information such as the data size, summary statistics for numerical variables, most common values for categorical variables, counts of missing values, and counts of unknown/undefined values. This file serves as a comprehensive report summarizing the findings of the data exploration analysis.

Overall, the code provides useful insights into the dataset's characteristics, which can inform further analysis and decision-making based on the data.

B.3 Code Output

data_exploration.txt

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

# Summary Statistics
          BMI  GenHlth  MentHlth  PhysHlth    Age  Education
              Income
count  253680   253680    253680    253680 253680     253680
    253680
mean    28.38    2.511     3.185     4.242  8.032       5.05
    6.054
```

```
std     6.609     1.068      7.413      8.718  3.054      0.9858
    2.071
min        12         1          0          0      1           1
         1
25%        24         2          0          0      6           4
         5
50%        27         2          0          0      8           5
         7
75%        31         3          2          3     10           6
         8
max        98         5         30         30     13           6
         8

# Categorical Variables

# 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
dtype: int64

# Unknown/undefined values
Diabetes_binary: Number of unknown/undefined values: 0
HighBP: Number of unknown/undefined values: 0
HighChol: Number of unknown/undefined values: 0
CholCheck: Number of unknown/undefined values: 0
BMI: Number of unknown/undefined values: 0
Smoker: Number of unknown/undefined values: 0
Stroke: Number of unknown/undefined values: 0
HeartDiseaseorAttack: Number of unknown/undefined values: 0
```

```
PhysActivity: Number of unknown/undefined values: 0
Fruits: Number of unknown/undefined values: 0
Veggies: Number of unknown/undefined values: 0
HvyAlcoholConsump: Number of unknown/undefined values: 0
AnyHealthcare: Number of unknown/undefined values: 0
NoDocbcCost: Number of unknown/undefined values: 0
GenHlth: Number of unknown/undefined values: 0
MentHlth: Number of unknown/undefined values: 0
PhysHlth: Number of unknown/undefined values: 0
DiffWalk: Number of unknown/undefined values: 0
Sex: Number of unknown/undefined values: 0
Age: Number of unknown/undefined values: 0
Education: Number of unknown/undefined values: 0
Income: Number of unknown/undefined values: 0
```

C Data Analysis

C.1 Code

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

```
# IMPORT
import pandas as pd
import numpy as np
import pickle
import statsmodels.formula.api as smf

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

# DATASET PREPARATIONS
# No dataset preparations are needed.

# DESCRIPTIVE STATISTICS
## Table 0: "Descriptive statistics of main binary variables
    ↪ and BMI stratified by sex"
df_sex = df.groupby('Sex').agg(
{'Diabetes_binary': ['mean','std'],
'PhysActivity': ['mean','std'],
'BMI': ['mean', 'std']}).reset_index()

df_sex.columns = ['Sex', 'Mean Diabetes', 'Std Diabetes', 'Mean
    ↪  PhysActivity', 'Std PhysActivity', 'Mean BMI', 'Std BMI'
    ↪ ]
df_sex.Sex = df_sex.Sex.replace({0: 'Female', 1: 'Male'})
df_sex.set_index('Sex', inplace=True)  # setting Sex as the
    ↪ index, for better row labels
```

```
df_sex.to_pickle(’table_0.pkl’)

# PREPROCESSING
# Creating dummy variables for categorical variables.
df = pd.get_dummies(df, columns=[’Sex’, ’Age’, ’Education’, ’
    ↪ Income’])

# ANALYSIS
## Table 1: "Results of Multiple Linear Regression testing
    ↪ association between physical activity level and diabetes,
    ↪  adjusting by age, sex, and BMI"
formula = ’Diabetes_binary ~ PhysActivity + BMI + Sex_1’
model = smf.ols(formula, data=df)
res = model.fit()
df1 = pd.DataFrame({
’coef’: res.params,
’p-value’: res.pvalues,
’conf_int_low’: res.conf_int().iloc[:, 0],
’conf_int_high’: res.conf_int().iloc[:, 1]
})
df1.to_pickle(’table_1.pkl’)

# SAVE ADDITIONAL RESULTS
additional_results = {
’Total number of observations’: df.shape[0],
’R-squared of the model’: res.rsquared,
}
with open(’additional_results.pkl’, ’wb’) as f:
    pickle.dump(additional_results, f)
```

C.2 Code Description

The provided code performs data analysis on the given dataset, which contains diabetes-related factors extracted from the CDC’s Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015.

The code starts by loading the dataset into a pandas DataFrame. Since the dataset is already preprocessed and contains no missing values, no further dataset preparations are needed.

Next, the code computes descriptive statistics on binary variables (Diabetes, Physical Activity) and the Body Mass Index (BMI) stratified by sex. This information is stored in a new DataFrame named ’df_sex’. The ’df_sex’ DataFrame is then saved as a pickle file named ’table_0.pkl’.

Following that, the code performs preprocessing by creating dummy variables for categorical variables (Sex, Age, Education, and Income).

The main analysis is done using a multiple linear regression model. The code builds a model using the formula 'Diabetes_binary ˜ PhysActivity + BMI + Sex_1', where PhysActivity represents the physical activity level, BMI represents the Body Mass Index, and Sex_1 represents the dummy variable for male sex. The model is fitted using the statsmodels library, and the results including coefficients, p-values, and confidence intervals are stored in a DataFrame named 'df1'. This DataFrame is saved as a pickle file named 'table_1.pkl'.

Additionally, the code saves some additional results in a pickle file named 'additional_results.pkl'. These additional results include the total number of observations in the dataset and the R-squared value of the fitted regression model.

C.3 Additional Results

The 'additional_results.pkl' file contains the following information:

1. Total number of observations: This is the total number of rows in the dataset, which represents the number of individuals included in the analysis.

2. R-squared of the model: This value indicates the proportion of variance in the dependent variable (Diabetes_binary) that can be explained by the independent variables (PhysActivity, BMI, and Sex_1) in the multiple linear regression model. A higher R-squared value indicates a better fit of the model to the data.

These additional results provide further insights into the analysis and can be used to interpret and evaluate the findings of the study.

C.4 Code Output

table_0.pkl

```
                Mean Diabetes  Std Diabetes  Mean PhysActivity  Std
       PhysActivity  Mean BMI  Std BMI
Sex
Female          0.1297         0.336              0.7442
              0.4363      28.13    7.088
Male            0.1516        0.3586              0.7723
              0.4194       28.7    5.928
```

table_1.pkl

```
                   coef    p-value  conf_int_low  conf_int_high
Intercept       -0.1155  3.15e-254       -0.1221        -0.1088
Sex_1[T.True]   0.01787   4.37e-40       0.01523        0.02052
PhysActivity   -0.07191          0        -0.075       -0.06882
BMI             0.01062          0       0.01042        0.01082
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'R-squared of the model': 0.05527          ,
}
```

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, Tuple, Optional, Any

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

# PREPARATION FOR ALL TABLES
# Define a shared mapping for labels that are common to all
   ↪ tables
shared_mapping: AbbrToNameDef = {
    'BMI': ('Body Mass Index', 'Measure of body fat based on
       ↪ height and weight that applies to adult men and women
       ↪ '),
    'Sex_1': ('Gender', 'Categorical variable, 1=Male, 0=Female
       ↪ ')
}

# 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))
mapping0 |= {
    'Mean Diabetes': ('Mean Diabetes', None),
    'Std Diabetes': ('Std. Diabetes', None),
    'Mean PhysActivity': ('Mean Phys. Activity', None),
    'Std PhysActivity': ('Std. Phys. Activity', None),
    'Mean BMI': ('Mean BMI', None),
    'Std BMI': ('Std. BMI', None),
}

abbrs_to_names0, legend0 = split_mapping(mapping0)
```

```
df0 = df0.rename(columns=abbrs_to_names0, index=abbrs_to_names0
    ↪ )

# Transpose the table to make it narrower
df0 = df0.T

# SAVE AS LATEX:
to_latex_with_note(
    df0, 'table_0.tex',
    caption="Descriptive Statistics of Main Binary Variables
        ↪ and Body Mass Index Stratified by Gender",
    label='table:DescriptiveStatistics',
    note="Descriptive statistics include mean and standard
        ↪ deviation of diabetes measure, physical activity, and
        ↪  body mass index.",
    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))
mapping1 |= {
    'coef': ('Coefficient', None),
    'p-value': ('P-value', None),
    'conf_int_low': ('CI (Lower)', None),
    'conf_int_high': ('CI (Higher)', None),
    'PhysActivity': ('Phys. Activity', 'Physical activity in
        ↪ the past 30 days'),
    'Intercept': ('Intercept', None),
    'Sex_1[T.True]': ('Gender Male', 'Categorical variable,
        ↪ Male'),
}

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='Multiple Linear Regression for Testing Association
        ↪  between Physical Activity Level and Diabetes,
        ↪ Adjusted by Age, Gender, and Body Mass Index',
    label='table:LinearRegression',
    note='Table includes the results of a multiple linear
        ↪ regression model on the relationship between diabetes
        ↪  and physical activity, adjusted by gender, age, and
        ↪ body mass index.',
    legend=legend1)
```

D.2 Provided Code

The code above is using the following provided functions:

```
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 Main Binary Variables and
   Body Mass Index Stratified by Gender}
\label{table:DescriptiveStatistics}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrr}
\toprule
Sex & Female & Male \\
\midrule
\textbf{Mean Diabetes} & 0.13 & 0.152 \\
\textbf{Std. Diabetes} & 0.336 & 0.359 \\
\textbf{Mean Phys. Activity} & 0.744 & 0.772 \\
\textbf{Std. Phys. Activity} & 0.436 & 0.419 \\
\textbf{Mean BMI} & 28.1 & 28.7 \\
\textbf{Std. BMI} & 7.09 & 5.93 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Descriptive statistics include mean and standard
   deviation of diabetes measure, physical activity, and body
   mass index.
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
% This latex table was generated from: ‘table_1.pkl‘
\begin{table}[h]
\caption{Multiple Linear Regression for Testing Association
   between Physical Activity Level and Diabetes, Adjusted by
   Age, Gender, and Body Mass Index}
\label{table:LinearRegression}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrlrr}
\toprule
 & Coefficient & P-value & CI (Lower) & CI (Higher) \\
\midrule
\textbf{Intercept} & -0.115 & $<$1e-06 & -0.122 & -0.109 \\
\textbf{Gender Male} & 0.0179 & $<$1e-06 & 0.0152 & 0.0205 \\
\textbf{Phys. Activity} & -0.0719 & $<$1e-06 & -0.075 & -0.0688
    \\
\textbf{Body Mass Index} & 0.0106 & $<$1e-06 & 0.0104 & 0.0108
   \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Table includes the results of a multiple linear
   regression model on the relationship between diabetes and
   physical activity, adjusted by gender, age, and body mass
   index.
\item \textbf{Body Mass Index}: Measure of body fat based on
   height and weight that applies to adult men and women
\item \textbf{Phys. Activity}: Physical activity in the past 30
    days
\item \textbf{Gender Male}: Categorical variable, Male
\end{tablenotes}
\end{threeparttable}
\end{table}
```