Unraveling the Combined Impact of Lifestyle and Obesity on Diabetes Risk

data-to-paper

April 18, 2024

Abstract

The escalating prevalence of diabetes is closely intertwined with lifestyle choices and the obesity epidemic, making the disentanglement of their contributions to disease onset a public health imperative. This investigation assesses the relative influence of dietary behaviors, physical activity, and body mass index (BMI) on diabetes prevalence among U.S. adults. Drawing on self-reported data from the Behavioral Risk Factor Surveillance System comprising over 250,000 participants in 2015, logistic regression analysis was utilized to elucidate complex relationships. The study identifies an inverse correlation between physical activity and fruit and vegetable intake with diabetes occurrence, while higher BMI is positively correlated. Furthermore, an interaction between BMI and physical activity suggests a diminishing protective effect of exercise on diabetes risk with increasing BMI. Demographic variables, including age, sex, and education, were also significant modulators. Despite the limitations inherent in the cross-sectional design and self-reporting, these findings advocate for the refinement of public health policies to curtail diabetes risk through multifaceted lifestyle interventions. Prospective longitudinal studies are warranted to verify causality and enhance the precision of diabetes prevention strategies.

Introduction

Diabetes, a chronic health condition characterized by the body's inability to regulate blood sugar either due to insufficient insulin production or impaired insulin sensitivity, continues to be a prevalent health issue globally[1]. Especially troubling is the escalating prevalence of diabetes, which is, in part, intertwined with the escalating obesity epidemic and pervasive unhealthy lifestyle choices[2]. The intertwined nature of these issues posits an urgent

need to study the collective influence of obesity and lifestyle choices on diabetes for the refinement of public health strategies.

Extensive research has been conducted on the effects of lifestyle factors, such as physical activity and dietary behaviors, and their influence on managing diabetes. In general, an inverse relationship has been observed between the incidence rate of diabetes and factors such as the intake of fruits and vegetables and the amount of physical activity performed[3, 4].Conversely, a direct relationship between increased Body Mass Index (BMI) and diabetes incidence has been established, underlining the critical role played by obesity in diabetes risk[5]. Despite these widely accepted relationships, some challenges remain. Although numerous studies have acknowledged the complex interplay of these lifestyle factors, obesity and diabetes, a comprehensive investigation considering a multitude of diabetes-influencing lifestyle factors concurrently in relation to obesity remains to be fully considered.

To shed light on this underexplored area, we have employed a dataset comprising of over 250,000 participants from the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015[6]. This comprehensive data source, encompassing various demographic, lifestyle, and obesity-related health indicators, provides rich insights into the multifaceted nature of diabetes risk. In addition to enabling an exploration of the individual effect of factors such as physical activity and diet, this dataset also provides us an opportunity to understand their combined effects along with BMI on diabetes prevalence.

We conducted a logistic regression analysis on the data using a methodological approach informed by established statistical methodologies[7, 8]. Our methods included evaluating relationships between various lifestyle factors, BMI, and diabetes risk, while controlling for age, sex, and education level. Further, we also tested for potential interaction effects, such as the influence of BMI on the relationship between physical activity and diabetes risk. The analysis carried out in this study is intended to offer a comprehensive understanding of the intertwined effects of lifestyle factors and obesity on diabetes prevalence.

Results

First, to understand the effects of lifestyle factors and high blood pressure on diabetes, we conducted a logistic regression analysis incorporating behavioral, physiological, and demographic variables. The initial model in Table

1 investigated the associations between diabetes occurrence and physical activity, fruit and vegetable consumption, BMI, high blood pressure, age, sex, and education level, where significant predictors of diabetes are identified. Physical activity reduced the odds of having diabetes with an odds ratio of 0.7233 ($p<10^{-6}$). Similarly, individuals who consumed fruits were less likely to report diabetes with odds ratios of 0.8976 and those who consumed vegetables also presented with reduced odds with an odds ratio of 0.879 (both $p<10^{-6}$). On the other hand, a unit increase in BMI was associated with higher odds of diabetes with an odds ratio of 1.089 ($p<10^{-6}$). High blood pressure was also included in the model, presenting a statistically significant coefficient. The model confirmed the constant term was significant ($p<10^{-6}$), providing an estimated baseline odds for the reference category of each variable. In this full model, it has an Akaike Information Criterion (AIC) of 1.81 10^5, assessing the relative fit of the model to the given dataset.

Table 1: Associations between physical activity, fruit and vegetable consumption, BMI, age, sex and education level with diabetes

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Constant	-4.89	0.0512	-95.5	$<10^{-6}$	-4.99	-4.79
Physical Activity	-0.324	0.0134	-24.3	$<10^{-6}$	-0.35	-0.298
Fruit Consumption	-0.108	0.013	-8.29	$<10^{-6}$	-0.133	-0.0824
Vegetable Consumption	-0.129	0.0151	-8.52	$<10^{-6}$	-0.158	-0.0991
BMI	0.0851	0.000878	96.9	$<10^{-6}$	0.0833	0.0868
Age Category	0.218	0.00239	91	$<10^{-6}$	0.213	0.222
Sex	0.246	0.0123	20.1	$<10^{-6}$	0.222	0.27
Education Level	-0.214	0.00599	-35.7	$<10^{-6}$	-0.226	-0.202

The model coefficients, standard errors, z-scores, p-values, and 95% confidence intervals are reported for each variable in the logistic regression model.
Sex: 0: Female, 1: Male
Age Category: 13-level age category in intervals of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 - 79, 13 = 80 or older)
Education Level: 1=Never attended school, 2=Elementary, 3=Some high school, 4=High school, 5=Some college, 6=College
BMI: Body Mass Index
Physical Activity: Physical Activity in past 30 days (0 = no, 1 = yes)
Fruit Consumption: Consume one fruit or more each day (0 = no, 1 = yes)
Vegetable Consumption: Consume one vegetable or more each day (0 = no, 1 = yes)
z: Z-score for the hypothesis test of zero Coefficient

Then, to test whether BMI modifies the association between physical activity and diabetes, we analyzed an interaction model presented in Table

2. The interaction between physical activity and BMI was statistically significant with a coefficient of 0.0114 ($p<10^{-6}$), indicating that the protective effect of physical activity on the risk of diabetes is influenced by the individual's BMI. The positive sign of the interaction coefficient suggests this protective effect becomes less prominent as BMI increases. The constants remained statistically significant ($p<10^{-6}$), and the AIC for Model 2 was 1.809 10^5, indicating a slight improvement in model fit over Model 1.

Table 2: Effect modification by BMI on the association between physical activity and diabetes

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Constant	-4.68	0.0602	-77.8	$<10^{-6}$	-4.8	-4.56
Physical Activity	-0.677	0.0557	-12.2	$<10^{-6}$	-0.786	-0.568
BMI	0.0781	0.00137	57	$<10^{-6}$	0.0754	0.0808
Physical Activity * BMI	0.0114	0.00174	6.53	$<10^{-6}$	0.00797	0.0148
Fruit Consumption	-0.107	0.013	-8.23	$<10^{-6}$	-0.132	-0.0816
Vegetable Consumption	-0.128	0.0151	-8.48	$<10^{-6}$	-0.158	-0.0984
Age Category	0.218	0.00239	91.1	$<10^{-6}$	0.213	0.223
Sex	0.244	0.0123	19.9	$<10^{-6}$	0.22	0.268
Education Level	-0.212	0.006	-35.4	$<10^{-6}$	-0.224	-0.2

The model coefficients, standard errors, z-scores, p-values, and 95% confidence intervals are reported for each variable in the logistic regression model.
Sex: 0: Female, 1: Male
Age Category: 13-level age category in intervals of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 - 79, 13 = 80 or older)
Education Level: 1=Never attended school, 2=Elementary, 3=Some high school, 4=High school, 5=Some college, 6=College
BMI: Body Mass Index
Physical Activity: Physical Activity in past 30 days (0 = no, 1 = yes)
Fruit Consumption: Consume one fruit or more each day (0 = no, 1 = yes)
Vegetable Consumption: Consume one vegetable or more each day (0 = no, 1 = yes)
z: Z-score for the hypothesis test of zero Coefficient
Physical Activity * BMI: Interaction term between Physical Activity and Body Mass Index

Finally, the overall dataset's impact and reliability of the logistic regression models were supported by a considerable sample size of 253,680 observations. This robust dataset from a wide cross-section of the U.S. population provides a powerful basis for the analyses, enhancing the generalizability of the study conclusions.

In summary, these results suggest that engagement in physical activities and healthy dietary habits are negatively associated with the likelihood

of having diabetes, with significant effect sizes observed for both physical activity and fruit and vegetable consumption. Conversely, higher BMI is positively associated with an increased risk of diabetes, with the observed interaction effect with physical activity highlighting the multifaceted nature of diabetes risk factors. Furthermore, the results affirm the significance of demographic factors, including age, sex, and education level, in the context of diabetes prevalence.

Discussion

Our study aimed to explore the intricate intertwining relationship of lifestyle choices, obesity, and the risk of diabetes–an escalating global health crisis. This undertaking was motivated by the cumulative understanding that diabetes prevalence is not just tightly interlaced with obesity but also highly influenced by lifestyle behaviors, a notion consistently borne out in prior research such as those conducted by Bellou et al.[1] and Powell-Wiley et al.[2].

Using the 2015 BRFSS dataset and logistic regression, we investigated associations between diabetes and variables encompassing lifestyle aspects and obesity. The study's findings indicate an inverse association of physical activity and fruit and vegetable intake on diabetes. This aligns with studies like those by Reis et al.[3] and Lv et al.[9] that highlighted similar inverse associations with diabetes. Notably, though, the relative influence of these variables on diabetes demonstrated in our study is more pronounced. The positive correlation between BMI and diabetes in our study mimics the findings from a broad range of studies such as Hjerkind et al.[10] and Lv et al.[9], reaffirming the pivotal role of obesity in diabetes risk. An interaction effect between physical activity and BMI was also observed, suggesting a diminishing protective influence of physical activity on diabetes with increasing BMI.

While these implications are compelling, the study's design and execution warrant some considerations. The use of a cross-sectional design limits our ability to discern definitive causal relationships. Additionally, the reliance on self-reported data may introduce deviations as participants might over or understate lifestyle factors like activity level or diet, potentially influencing the interpretation of analyses. Lastly, demographic factors such as age, sex, and education also exhibited notable associations with diabetes prevalence, replicating prior research implications such as Zhang et al.[11] and Sung et al.[12], further demonstrating the multifactorial nature of dia-

betes risk.

Despite these limitations, this study makes significant strides towards understanding the mutual influences of lifestyle factors and obesity on diabetes. These valuable insights uphold the link between healthier lifestyles–physical activity and fruit and vegetable-rich diets–and lowered diabetes risk, with attenuated effectiveness with increasing BMI. These findings enrich the existing body of knowledge and emphasize the need for multifaceted interventions in managing diabetes risk.

Prospective avenues of research could look at replicating these findings across other data sources and incorporating a broader array of lifestyle and demographic factors. The implementation of prospective longitudinal study designs could further reinforce causal inferences and thereby aid in refining diabetes prevention strategies. The findings of this study underscore the importance of integrative insights in developing public health practices and strategies, both at an individual and population level, to more effectively manage and combat the challenge posed by diabetes.

Methods

Data Source

The dataset utilized for this study comprised diabetes-related health indicators derived from the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. This dataset is a compilation of over 250,000 responses from U.S. adults, who participated in the annual health-related telephone survey conducted by the CDC. The survey collects information on health-related risk behaviors, chronic health conditions, and the utilization of preventive services.

Data Preprocessing

The dataset was pre-existing in a format requiring no additional preprocessing for the variables pertinent to our study. The variables of interest for our analysis—the binary indicators for diabetes, physical activity, fruit and vegetable consumption, as well as numerical ones such as body mass index (BMI), age, gender, and level of education—were already encoded adequately for logistic regression analysis. Consequently, no further preprocessing steps were applied to this dataset within our analytical pipeline.

Data Analysis

To explore the relationships among lifestyle factors, obesity, and the likelihood of being diagnosed with diabetes, we conducted logistic regression analysis. Initially, we evaluated the associations between physical activity, diet, BMI, and diabetes while controlling for age, sex, and education. Subsequently, to assess the effect modification by BMI on the physical activity-diabetes association, a similar logistic regression model incorporating an interaction term between physical activity and BMI was also formulated. Both models employed a method to adjust for all possible confounders in the analysis. The AIC statistic was collected for each of the models as a measure of model fit. The analysis provided results highlighting relationships between the variables of interest and the occurrence of diabetes, accounting for the specified demographic covariates.

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] V. Bellou, L. Belbasis, I. Tzoulaki, and E. Evangelou. Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses. *PLoS ONE*, 13, 2018.

[2] T. Powell-Wiley, P. Poirier, L. Burke, J. Despres, P. Gordon-Larsen, C. Lavie, S. Lear, C. Ndumele, I. Neeland, P. Sanders, and M. St-Onge. Obesity and cardiovascular disease: A scientific statement from the american heart association. *Circulation*, 143:e984 – e1010, 2021.

[3] J. Reis, C. Loria, P. Sorlie, Yikyung Park, A. Hollenbeck, and A. Schatzkin. Lifestyle factors and risk for new-onset diabetes. *Annals of Internal Medicine*, 155:292 – 299, 2011.

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

[5] T. Schnurr, Hermina Jakupovi, Germn D. Carrasquilla, L. ngquist, N. Grarup, T. Srensen, A. Tjnneland, K. Overvad, O. Pedersen,

T. Hansen, and T. Kilpelinen. Obesity, unfavourable lifestyle and genetic risk of type 2 diabetes: a case-cohort study. *Diabetologia*, 63:1324–1332, 2020.

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

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

[8] M. Knol, I. van der Tweel, D. Grobbee, M. Numans, and M. Geerlings. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *International journal of epidemiology*, 36 5:1111–8, 2007.

[9] J. Lv, Canqing Yu, Yu Guo, Z. Bian, Ling Yang, Yiping Chen, Ximin Hu, W. Hou, Junshi Chen, Zhengming Chen, L. Qi, and Liming Li. Adherence to a healthy lifestyle and the risk of type 2 diabetes in chinese adults. *International Journal of Epidemiology*, 46:1410 – 1420, 2017.

[10] K. V. Hjerkind, J. 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.

[11] Cuilin Zhang, Deirdre K. Tobias, J. Chavarro, W. Bao, Dong D. Wang, S. Ley, and F. Hu. Adherence to healthy lifestyle and risk of gestational diabetes mellitus: prospective cohort study. *The BMJ*, 349, 2014.

[12] K. Sung, W. Jeong, S. Wild, and C. Byrne. Combined influence of insulin resistance, overweight/obesity, and fatty liver as risk factors for type 2 diabetes. *Diabetes Care*, 35:717 – 722, 2012.

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

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

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

  # Print the size of the Data
  f.write("# Data Size\n")
  f.write(f"Number of Rows: {df.shape[0]}\n")
  f.write(f"Number of Columns: {df.shape[1]}\n\n")
```

```
  # Print summary statistics
  f.write("# Summary Statistics\n")
  summary_stats = df.describe(include='all').transpose()
  f.write(f"{summary_stats}\n\n")

  # Identify & count the categorical variables
  f.write("# Categorical Variables\n")
  categorical_columns = df.select_dtypes(include=['bool', '
      ↪ object']).columns
  if len(categorical_columns) == 0:
    f.write("Not applicable\n\n")
  else:
    categorical_variable_stats = df[categorical_columns].
        ↪ describe().transpose()
    f.write(f"{categorical_variable_stats}\n\n")

  # Print counts of missing/undefined values
  f.write("# Missing Values\n")
  missing_values = df.isnull().sum().to_frame(name='Missing
      ↪ values count')
  if missing_values.empty:
    f.write("Not applicable\n\n")
  else:
    f.write(f"{missing_values}\n\n")

  # Anything else you deem relevant.
  # For example, you might want to count each classes in
      ↪ Diabetes_binary
  f.write("# Count of classes in Diabetes_binary\n")
  diabetes_binary_counts = df['Diabetes_binary'].value_counts()
      ↪ .to_frame(name='Counts')
  f.write(f"{diabetes_binary_counts}\n\n")

f.close()
```

B.2 Code Description

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

First, the code loads the dataset into a pandas DataFrame.

The code then writes various analysis results into a text file named "data_exploration.txt".

The code starts by printing the size of the dataset, including the number of rows and columns.

Next, the code calculates and writes summary statistics for all the columns in the dataset. This includes measures such as count, mean, standard deviation, minimum, quartiles, and maximum for numerical columns, as well as count, unique, top, and frequency for categorical columns.

The code identifies and counts the categorical variables in the dataset, and writes information about their counts and unique values.

The code then checks for missing values in the dataset and writes the count of missing values for each column.

Finally, the code performs an additional analysis step where it counts the number of occurrences for each class in the "Diabetes_binary" column, indicating the frequency of positive and negative instances for diabetes in the dataset.

All the analysis results are written into the "data_exploration.txt" file, providing valuable insights about the dataset's size, summary statistics, categorical variables, missing values, and the distribution of the diabetes class.

This data exploration process allows researchers to gain a better understanding of the dataset, identify potential data quality issues, and make informed decisions for subsequent data preprocessing and modeling steps.

B.3 Code Output

data_exploration.txt

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

# Summary Statistics
                      count    mean     std  min  25%  50%  75%
                  max
Diabetes_binary      253680  0.1393  0.3463    0    0    0    0
      1
HighBP               253680   0.429  0.4949    0    0    0    1
      1
HighChol             253680  0.4241  0.4942    0    0    0    1
      1
CholCheck            253680  0.9627  0.1896    0    1    1    1
      1
BMI                  253680   28.38   6.609   12   24   27   31
     98
Smoker               253680  0.4432  0.4968    0    0    0    1
      1
Stroke               253680 0.04057  0.1973    0    0    0    0
      1
```

```
HeartDiseaseorAttack 253680 0.09419 0.2921    0    0    0    0
       1
PhysActivity         253680  0.7565 0.4292    0    1    1    1
       1
Fruits               253680  0.6343 0.4816    0    0    1    1
       1
Veggies              253680  0.8114 0.3912    0    1    1    1
       1
HvyAlcoholConsump    253680  0.0562 0.2303    0    0    0    0
       1
AnyHealthcare        253680  0.9511 0.2158    0    1    1    1
       1
NoDocbcCost          253680 0.08418 0.2777    0    0    0    0
       1
GenHlth              253680   2.511  1.068    1    2    2    3
       5
MentHlth             253680   3.185  7.413    0    0    0    2
      30
PhysHlth             253680   4.242  8.718    0    0    0    3
      30
DiffWalk             253680  0.1682 0.3741    0    0    0    0
       1
Sex                  253680  0.4403 0.4964    0    0    0    1
       1
Age                  253680   8.032  3.054    1    6    8   10
      13
Education            253680    5.05 0.9858    1    4    5    6
       6
Income               253680   6.054  2.071    1    5    7    8
       8

# Categorical Variables
Not applicable

# Missing Values
                     Missing values count
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

# Count of classes in Diabetes_binary
                  Counts
Diabetes_binary
0                 218334
1                  35346
```

C Data Analysis

C.1 Code

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

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

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

# DATASET PREPARATIONS
# No dataset preparations are needed

# DESCRIPTIVE STATISTICS
# No descriptive statistics table is needed

# PREPROCESSING
# No preprocessing is needed, because our variables of interest
    ↪  are already binary or numerical

# ANALYSIS
## Table 1: "Associations between physical activity, diet, BMI
    ↪ and diabetes"
model = logit('Diabetes_binary ~ PhysActivity + Fruits +
    ↪ Veggies + BMI + Age + Sex + Education',
    data=add_constant(df)).fit()
```

```
df1 = model.summary2().tables[1]
df1.to_pickle('table_1.pkl')

## Table 2: "Effect modification by BMI on the association
    ↪ between physical activity and diabetes"
model_interaction = logit('Diabetes_binary ~ PhysActivity*BMI +
    ↪  Fruits + Veggies + Age + Sex + Education',
    data=add_constant(df)).fit()
df2 = model_interaction.summary2().tables[1]
df2.to_pickle('table_2.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
    'Total number of observations': len(df),
    'Model 1 AIC': model.aic,
    'Model 2 AIC': model_interaction.aic,
}
with open('additional_results.pkl', 'wb') as f:
    pickle.dump(additional_results, f)
```

C.2 Code Description

The code performs data analysis on the diabetes-related factors dataset obtained from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The goal of the analysis is to examine the associations between physical activity, diet, body mass index (BMI), and diabetes, as well as to explore the effect modification of BMI on the association between physical activity and diabetes.

The code first loads the dataset into a pandas dataframe. The dataset is already clean and does not contain any missing values.

Next, the code fits two logistic regression models to examine the associations between the variables of interest and diabetes. The first model, referred to as "Table 1", includes the predictors PhysActivity (physical activity), Fruits, Veggies, BMI, Age, Sex, and Education. The model is fit using the logit function from the statsmodels library, after adding a constant term to the dataframe using the add_constant function. The summary results of the model, including coefficient estimates, standard errors, p-values, and other statistics, are saved in a pickle file named "table_1.pkl".

The second model, referred to as "Table 2", examines the effect modification of BMI on the association between physical activity and diabetes. In addition to the predictors included in Table 1, this model also includes an interaction term between PhysActivity and BMI. Similar to Table 1, the model is fit using the logit function and the results are saved in a pickle file

named "table_2.pkl".

Finally, the code saves additional results in a pickle file named "additional_results.pkl". These results include the total number of observations (i.e., the number of rows in the dataset) and the AIC (Akaike Information Criterion) values for both Model 1 and Model 2. AIC is a measure of the goodness of fit of the model, with lower values indicating a better fit.

The purpose of saving the results in pickle files is to allow for easy access and retrieval of the analysis results for further examination or reporting purposes.

C.3 Code Output

table_1.pkl

```
               Coef.   Std.Err.       z     P>|z|   [0.025
                  0.975]
Intercept     -4.891    0.05122 -95.48         0   -4.991
    -4.79
PhysActivity -0.3242    0.01337 -24.25  6.59e-130 -0.3505
    -0.298
Fruits        -0.1079      0.013 -8.295   1.09e-16 -0.1333
    -0.08237
Veggies       -0.1287     0.0151 -8.518   1.63e-17 -0.1583
    -0.09905
BMI           0.08506 0.0008776  96.93          0 0.08334
    0.08678
Age            0.2177  0.002392     91          0   0.213
    0.2224
Sex             0.246   0.01226  20.06   1.62e-89   0.222
    0.27
Education    -0.2139   0.00599 -35.71  3.08e-279 -0.2256
    -0.2021
```

table_2.pkl

```
                   Coef. Std.Err.       z      P>|z|    [0.025
                      0.975]
Intercept          -4.681  0.06017 -77.79          0    -4.799
    -4.563
PhysActivity      -0.6773  0.05565 -12.17   4.49e-34   -0.7864
    -0.5682
BMI               0.07811  0.00137  57.01          0   0.07543
    0.0808
PhysActivity:BMI 0.01138 0.001744  6.528   6.66e-11 0.007965
    0.0148
Fruits             -0.107    0.013 -8.235    1.8e-16   -0.1325
    -0.08155
```

```
Veggies            -0.128   0.01509  -8.482    2.22e-17   -0.1576
    -0.09843
Age                0.2179  0.002393   91.06           0    0.2132
    0.2226
Sex                0.2437   0.01226   19.87    7.97e-88    0.2196
    0.2677
Education          -0.212  0.005996  -35.36   7.36e-274   -0.2238
    -0.2003
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'Model 1 AIC': 1.81e+05          ,
    'Model 2 AIC': 1.809e+05         ,
}
```

D LaTeX Table Design

D.1 Code

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

```python
# IMPORT
import pandas as pd
from my_utils import to_latex_with_note, is_str_in_df,
    ↪ split_mapping, AbbrToNameDef

# PREPARATION FOR ALL TABLES
shared_mapping: AbbrToNameDef = {
    'Intercept': ('Constant', None),
    'Sex': ('Sex', '0: Female, 1: Male'),
    'Age': ('Age Category', '13-level age category in intervals
        ↪ of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 -
        ↪ 79, 13 = 80 or older)'),
    'Education': ('Education Level', '1=Never attended school,
        ↪ 2=Elementary, 3=Some high school, 4=High school, 5=
        ↪ Some college, 6=College'),
    'BMI': ('BMI', 'Body Mass Index'),
    'PhysActivity': ('Physical Activity', 'Physical Activity in
        ↪  past 30 days (0 = no, 1 = yes)'),
    'Fruits': ('Fruit Consumption', 'Consume one fruit or more
        ↪ each day (0 = no, 1 = yes)'),
    'Veggies': ('Vegetable Consumption', 'Consume one vegetable
        ↪  or more each day (0 = no, 1 = yes)'),
    'z': ('z', 'Z-score for the hypothesis test of zero
        ↪ Coefficient')
}
```

```
# 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="Associations between physical activity, fruit and
        ↪ vegetable consumption, BMI, age, sex and education
        ↪ level with diabetes",
    label='table:associations_physical_activity_BMI_diabetes',
    note="The model coefficients, standard errors, z-scores, p-
        ↪ values, and 95% confidence intervals are reported for
        ↪  each variable in the logistic regression model.",
    legend=legend1)


# 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))
mapping2 |= {
    'PhysActivity:BMI': ('Physical Activity * BMI', '
        ↪ Interaction term between Physical Activity and Body
        ↪ Mass Index')
}
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="Effect modification by BMI on the association
        ↪ between physical activity and diabetes",
    label='table:effect_modification_physical_activity_diabetes
        ↪ ',
    note="The model coefficients, standard errors, z-scores, p-
        ↪ values, and 95% confidence intervals are reported for
        ↪  each variable in the logistic regression model.",
```

```
    legend=legend2)
```

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_1.tex

```
% This latex table was generated from: `table_1.pkl`
\begin{table}[h]
\caption{Associations between physical activity, fruit and
    vegetable consumption, BMI, age, sex and education level
    with diabetes}
```

```
\label{table:associations_physical_activity_BMI_diabetes}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err. & z & P$>$\textbar{}z\textbar{} & [0.025 &
    0.975] \\
\midrule
\textbf{Constant} & -4.89 & 0.0512 & -95.5 & $<$1e-06 & -4.99 &
    -4.79 \\
\textbf{Physical Activity} & -0.324 & 0.0134 & -24.3 & $<$1e-06
    & -0.35 & -0.298 \\
\textbf{Fruit Consumption} & -0.108 & 0.013 & -8.29 & $<$1e-06
   & -0.133 & -0.0824 \\
\textbf{Vegetable Consumption} & -0.129 & 0.0151 & -8.52 & $<
   $1e-06 & -0.158 & -0.0991 \\
\textbf{BMI} & 0.0851 & 0.000878 & 96.9 & $<$1e-06 & 0.0833 &
   0.0868 \\
\textbf{Age Category} & 0.218 & 0.00239 & 91 & $<$1e-06 & 0.213
    & 0.222 \\
\textbf{Sex} & 0.246 & 0.0123 & 20.1 & $<$1e-06 & 0.222 & 0.27
   \\
\textbf{Education Level} & -0.214 & 0.00599 & -35.7 & $<$1e-06
   & -0.226 & -0.202 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item The model coefficients, standard errors, z-scores, p-
   values, and 95\% confidence intervals are reported for each
    variable in the logistic regression model.
\item \textbf{Sex}: 0: Female, 1: Male
\item \textbf{Age Category}: 13-level age category in intervals
    of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 - 79, 13 =
    80 or older)
\item \textbf{Education Level}: 1=Never attended school, 2=
   Elementary, 3=Some high school, 4=High school, 5=Some
   college, 6=College
\item \textbf{BMI}: Body Mass Index
\item \textbf{Physical Activity}: Physical Activity in past 30
    days (0 = no, 1 = yes)
\item \textbf{Fruit Consumption}: Consume one fruit or more
   each day (0 = no, 1 = yes)
\item \textbf{Vegetable Consumption}: Consume one vegetable or
   more each day (0 = no, 1 = yes)
\item \textbf{z}: Z-score for the hypothesis test of zero
   Coefficient
\end{tablenotes}
```

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

table_2.tex

```
% This latex table was generated from: `table_2.pkl`
\begin{table}[h]
\caption{Effect modification by BMI on the association between
    physical activity and diabetes}
\label{table:effect_modification_physical_activity_diabetes}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err. & z & P$>$\textbar{}z\textbar{} & [0.025 &
     0.975] \\
\midrule
\textbf{Constant} & -4.68 & 0.0602 & -77.8 & $<$1e-06 & -4.8 &
    -4.56 \\
\textbf{Physical Activity} & -0.677 & 0.0557 & -12.2 & $<$1e-06
     & -0.786 & -0.568 \\
\textbf{BMI} & 0.0781 & 0.00137 & 57 & $<$1e-06 & 0.0754 &
    0.0808 \\
\textbf{Physical Activity * BMI} & 0.0114 & 0.00174 & 6.53 & $<
    $1e-06 & 0.00797 & 0.0148 \\
\textbf{Fruit Consumption} & -0.107 & 0.013 & -8.23 & $<$1e-06
    & -0.132 & -0.0816 \\
\textbf{Vegetable Consumption} & -0.128 & 0.0151 & -8.48 & $<
    $1e-06 & -0.158 & -0.0984 \\
\textbf{Age Category} & 0.218 & 0.00239 & 91.1 & $<$1e-06 &
    0.213 & 0.223 \\
\textbf{Sex} & 0.244 & 0.0123 & 19.9 & $<$1e-06 & 0.22 & 0.268
    \\
\textbf{Education Level} & -0.212 & 0.006 & -35.4 & $<$1e-06 &
    -0.224 & -0.2 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item The model coefficients, standard errors, z-scores, p-
    values, and 95\% confidence intervals are reported for each
     variable in the logistic regression model.
\item \textbf{Sex}: 0: Female, 1: Male
\item \textbf{Age Category}: 13-level age category in intervals
     of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 - 79, 13 =
     80 or older)
\item \textbf{Education Level}: 1=Never attended school, 2=
    Elementary, 3=Some high school, 4=High school, 5=Some
```

```
    college, 6=College
\item \textbf{BMI}: Body Mass Index
\item \textbf{Physical Activity}: Physical Activity in past 30
    days (0 = no, 1 = yes)
\item \textbf{Fruit Consumption}: Consume one fruit or more
    each day (0 = no, 1 = yes)
\item \textbf{Vegetable Consumption}: Consume one vegetable or
    more each day (0 = no, 1 = yes)
\item \textbf{z}: Z-score for the hypothesis test of zero
    Coefficient
\item \textbf{Physical Activity * BMI}: Interaction term
    between Physical Activity and Body Mass Index
\end{tablenotes}
\end{threeparttable}
\end{table}
```

E Calculation Notes

- exp(-0.324) = 0.7233

  Calculating odds ratio from logistic regression coefficient for physical activity

- exp(-0.108) = 0.8976

  Calculating odds ratio from logistic regression coefficient for fruit consumption

- exp(-0.129) = 0.879

  Calculating odds ratio from logistic regression coefficient for vegetable consumption

- exp(0.0851) = 1.089

  Calculating odds ratio from logistic regression coefficient for BMI