Interplay of Physical Activity, Diet, and BMI in Determining Diabetes Risk

data-to-paper

March 30, 2024

Abstract

The escalating rates of diabetes amidst growing obesity and sedentary lifestyles globally constitute a serious public health crisis. This study fills a research void by quantifying the relationships between modifiable lifestyle factors and diabetes risk, an area that has been insufficiently mapped in large-scale population data. We analyze the expansive 2015 Behavioral Risk Factor Surveillance System dataset, encompassing over a quarter-million U.S. adult respondents, to operationalize this linkage. Methodologically, we eschewed intricate statistical models in favor of logistic regression to distill clear associations between lifestyle behaviors – specifically physical activity and dietary patterns – body mass index (BMI), and the presence of diabetes. Our principal findings underscore a negative association of diabetes prevalence with physical activity, and fruit and vegetable intake, whereas a positive link is observed with BMI, augmenting the risk profile notably. Further, the interaction between physical activity and BMI emerges as critical, with higher BMI seen to attenuate the protective benefits of physical activity. Demographics such as greater age and male sex are additional risk amplifiers, while higher education attenuates risk. The analysis, while constrained by the self-reported and cross-sectional nature of the dataset, thus steering clear of causal conclusions, sets the cornerstone for future in-depth longitudinal studies. It also imparts a crucial impetus for public health strategies to prioritize modifiable lifestyle factors to alleviate disease burden.

Introduction

The global rise in diabetes, understood as both a direct and indirect consequence of obesity, underscores a major public health concern [1, 2]. Lifestyle factors, particularly dietary habits and physical activity, significantly influence metabolic health and insulin resistance, thereby implicating them as critical elements in diabetes onset and progression [3, 4, 5, 6]. Consequently, understanding the relationship between these modifiable factors and diabetes risk assumes immense importance.

While ample evidence exists elucidating the individual associations between these lifestyle factors and diabetes risk, an integrated and comprehensive analysis of their interdependence, differential contributions, and confluence on diabetes risk is less clear [7, 8, 9]. Specifically, it is unclear how the beneficial effects of physical activity may be modulated by body mass index (BMI), a measure of obesity, thereby necessitating a finer-grained exploration of the mechanistic interplay among these factors as it pertains to diabetes risk [10].

Our study bridges this knowledge gap by conducting a comprehensive analysis on the Behavioral Risk Factor Surveillance System (BRFSS) dataset, an extensive health-related survey aggregating data from over a quarter-million U.S. adults [11, 12]. The dataset's granularity offers rich insights into the lifestyle habits of a demographically diverse population, enabling an exploration of the correlates of diabetes risk in unprecedented depth and breadth [13, 14].

In our analysis, we employ logistic regression, a modeling approach explicitly suited to dichotomous outcomes such as diabetes diagnosis. It enables us to map the composite and interacting effects of lifestyle factors, BMI, and demographic variables altogether, providing a robust portrait of diabetes risk vis-à-vis these parameters [15, 16]. By elucidating the individual and intersecting implications of these factors, our study provides further insight into the multifaceted etiology of diabetes, specifically the protective role of physical activity and its complex interaction with BMI [17, 18].

Results

First, to explore the impact of lifestyle choices on diabetes, we utilized logistic regression models to examine the odds of having diabetes based on physical activity, diet, and BMI. As reported in Table 1, engaging in physical activity (-0.324), fruit (-0.108), and vegetable (-0.129) consumption were inversely associated with the presence of diabetes, suggesting decreased odds of having the condition by -0.2767, -0.1024, and -0.121 times, respectively, for individuals who engaged in these behaviors compared to those who did not. A higher BMI was associated with increased odds (0.0851), suggesting

0.08883 times increased odds of having diabetes per unit increase in BMI. Moreover, demographic factors such as age (0.218), sex (0.246), and education (-0.214) were also associated with diabetes risk, with older age and male sex increasing the odds, while higher education decreased them.

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, assessing how BMI may modify the relationship between physical activity and diabetes, a model including an interaction term was analyzed. The model, displayed in Table 2, indicates that physical activity's association with reduced diabetes risk becomes less pronounced as BMI increases. The interaction term's coefficient (0.0114) translates to a 1.14 percent increase in the effect of each BMI unit on diabetes odds for individuals who are physically active, underscoring the diminishing protective effect as BMI increases. The adjusted odds ratios for physical activity (-0.677) and BMI (0.0781) separately imply that physical activity is associated with lower odds of diabetes, while higher BMI is linked to higher odds, prior to considering their interaction.

The AIC values demonstrate that the inclusion of the interaction term

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

provides a marginally better fit to the data (1.81 10^5 for Model 1, 1.809 10^5 for Model 2), suggesting the interaction between physical activity and BMI is a relevant factor, though the difference is subtle.

In summary, results from the logistic regression models indicate associations between certain behaviors, demographics, and the likelihood of having diabetes. Physical activity and healthy dietary habits are associated with lower odds, whereas higher BMI is linked with increased odds of diabetes, displaying a nuanced interaction where the beneficial effects of physical activity on diabetes risk are attenuated at higher BMI levels. Age and male sex increase the odds, whereas higher educational attainment is linked with lower odds of diabetes, aligning with the expected refinement of risk based on demographic characteristics. Importantly, these associations do not imply causation and should be interpreted in the context of the cross-sectional study design.

Discussion

In our endeavor to contribute to the body of knowledge surrounding diabetes, a globally escalating health concern linked fundamentally to obesity and modifiable lifestyle factors [1, 2, 5], we dove into a comprehensive analysis of the BRFSS survey data [11, 12]. This allowed us to explore the nuanced interactions between lifestyle behaviors, BMI, and demographic factors in determining diabetes risk. Using the logistic regression approach [15, 16], we unravelled the multifactorial nature of diabetes etiology and the degrees to which each determinant affects an individual's likelihood of having diabetes.

Our results corroborate previous findings in the scientific literature [3, 6, 19] that point to the protective effect of physical activity and healthy dietary behaviors against diabetes. However, we provide an added layer of insight by illustrating how these benefits are moderated by BMI. Namely, our findings indicate a declining positive impact of physical activity on diabetes risk with increasing BMI levels, substantiating similar suggestions in existing research [10, 20]. Our study also strengthens the evidence base by reinforcing the roles of age, sex, and education level as important influences on diabetes risk [21, 22].

While our study provides foundational insights, it also offers avenues for further exploration given its limitations. The use of logistic regression, while beneficial for delineating clear associations, can oversimplify the intricate interrelationships between variables. There also exists a possibility of residual confounding due to immeasurable or unconsidered variables despite adjustment efforts. The assignment of dichotomy to the physical activity variable may fail to account for variability in the types and intensity of exercise, which could have differing impacts on diabetes risk. Inherent limitations in the dataset may affect the robustness of the findings. Given its cross-sectional design, it doesn't enable the establishing of causal relationships. Additionally, the self-reported nature of the data may introduce recall and reporting bias, potentially influencing the accuracy of respondents' details on lifestyle behaviors and medical history.

With these insightful findings, combined with acknowledged study limitations, we illuminate the need for a nuanced approach in public health strategies for diabetes prevention. Our results underscore that these strategies should not only promote physical activity and healthy dietary habits but also tackle obesity at its core. Furthermore, demographic factors like age, sex, and education level should form an integral part of the planning and execution of these health strategies, given their salient influence on diabetes risk. Lastly, our work reiterates the importance of future longitudinal research to delve deeper into the intertwined relationships between these risk factors, particularly to differentiate their relative contribution to diabetes onset and progression, and to establish causality. These directions for future research could provide comprehensive and integrated strategies for diabetes prevention, which would be indispensable in curbing this global health issue.

Methods

Data Source

The study utilized a dataset comprising diabetes-related factors extracted from a health-related telephone survey conducted annually. The survey source collates responses from a broad demographic, encompassing over 400,000 U.S. participants, to gather data on health behaviors, chronic health conditions, and usage of preventive services. The dataset for our analysis included cleaned and complete records of 253,680 respondents, spanning 22 diverse health indicators. These indicators encompassed binary markers of disease status and health behaviors, ordinal self-assessments of health, and numerical evaluations such as Body Mass Index.

Data Preprocessing

In our study, the dataset required no additional preprocessing prior to analysis. Given the dataset provided was already cleansed of missing values and appropriately formatted, we proceeded directly to analysis. The inherent structure of the dataset featured variables suitable for logistic regression analysis without necessitating further manipulation or transformation. As such, our methodological approach ensured an unaltered assessment of the data as furnished.

Data Analysis

For the analysis, we employed logistic regression, a statistical method for modeling dichotomous outcome variables, to explore associations between lifestyle behaviors, body mass index (BMI), demographic factors, and the risk of diabetes. A series of models were constructed to gain insights into both independent and interactive effects. Initially, the main effects of physical activity, dietary patterns (fruit and vegetable intake), BMI, age, sex, and education level on diabetes presence were examined. Subsequently, an interaction term was incorporated to discern modifications in the effect of physical activity on diabetes risk across different BMI levels. Demographics were included as covariates to adjust for potential confounders within the model. The evaluation of goodness-of-fit and comparison between models were facilitated by the examination of Akaike information criteria (AIC) values. The final extracted models provided an estimation of effects in the form of odds ratio, with accompanying confidence intervals and p-values, allowing for a comprehensive analysis of the contributing factors to diabetes risk.

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

[2] E. McAllister, N. Dhurandhar, S. Keith, L. Aronne, J. Barger, M. Baskin, R. Benca, J. Biggio, M. Boggiano, J. Eisenmann, M. Elobeid, K. Fontaine, P. Gluckman, Erin C. Hanlon, P. Katzmarzyk, A. Pietrobelli, D. Redden, D. Ruden, Chenxi Wang, R. Waterland, Suzanne M. Wright, and D. Allison. Ten putative contributors to the obesity epidemic. *Critical Reviews in Food Science and Nutrition*, 49:868 – 913, 2009.

[3] Yanping Li, Dong D. Wang, S. Ley, Malik Vasanti, A. Howard, Yu na He, and F. Hu. Time trends of dietary and lifestyle factors and their potential impact on diabetes burden in china. *Diabetes Care*, 40:1685 – 1694, 2017.

[4] R. Wing, W. Lang, T. Wadden, M. Safford, W. Knowler, A. Bertoni, James O Hill, F. Brancati, A. Peters, and L. Wagenknecht. Benefits of modest weight loss in improving cardiovascular risk factors in overweight and obese individuals with type 2 diabetes. *Diabetes Care*, 34:1481 – 1486, 2011.

[5] 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.

[6] 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.

[7] 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.

[8] S. van Oort, J. Beulens, A. V. van Ballegooijen, D. Grobbee, and S. Larsson. Association of cardiovascular risk factors and lifestyle behaviors with hypertension. *Hypertension*, 76:1971 – 1979, 2020.

[9] Zixin Zeng, Yuqian Bian, Y. Cui, Donghui Yang, Yafeng Wang, and Chuanhua Yu. Physical activity dimensions and its association with risk of diabetes in middle and older aged chinese people. *International Journal of Environmental Research and Public Health*, 17, 2020.

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

[12] Carol Pierannunzi, S. Hu, and L. Balluz. A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (brfss), 20042011. *BMC Medical Research Methodology*, 13:49 – 49, 2013.

[13] D. Nelson, D. Holtzman, Julie Bolen, C. Stanwyck, and K. Mack. Reliability and validity of measures from the behavioral risk factor surveillance system (brfss). *Sozial- und Praventivmedizin*, 46 Suppl 1:S3–42, 2001.

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

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

[16] P. Peduzzi, J. Concato, Elizabeth Kemper, T. Holford, and A. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49 12:1373–9, 1996.

[17] A. Harding, N. Wareham, S. Bingham, K. Khaw, R. Luben, A. Welch, and N. Forouhi. Plasma vitamin c level, fruit and vegetable consumption, and the risk of new-onset type 2 diabetes mellitus: the european prospective investigation of cancer–norfolk prospective study. *Archives of internal medicine*, 168 14:1493–9, 2008.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] W. Knowler, E. Barrett-Connor, S. Fowler, R. Hamman, J. Lachin, Elizabeth A. Walker, and David M. Nathan. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England journal of medicine*, 346 6:393–403, 2002.

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:

```
# 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) - 1 = -0.2767

  decreased odds of physical activity

- exp(-0.108) - 1 = -0.1024

  decreased odds of fruit consumption decreased odds of fruit consumption decreased odds of fruit consumption

- exp(-0.129) - 1 = -0.121

  decreased odds of vegetable consumption

- exp(0.0851) - 1 = 0.08883

  increased odds per BMI unit

- 0.0114 * 100 = 1.14

  percent increase