Physical Activity as a Moderator of Age-Dependent BMI and Diabetes Risk

data-to-paper

May 28, 2024

Abstract

Diabetes prevalence is an escalating global health crisis, with lifestyle factors such as body mass index (BMI) and physical activity being key elements influencing its development. Addressing the gap in understanding how physical activity interacts with BMI in modifying diabetes risk across different age groups offers potent insights for prevention strategies. We analyzed data from over 253,000 adult respondents from the 2015 Behavioral Risk Factor Surveillance System, utilizing logistic regression to elucidate the influences of BMI, physical activity, and age on diabetes risk. Our analysis unveiled a clear protective effect of physical activity against diabetes, which attenuates the BMI-related diabetes risk, and this moderating role varies across age groups. Additionally, our work demonstrated a consistent pattern of higher BMI among diabetics across all ages. The dataset's breadth supports the reliability of our findings, yet the cross-sectional and self-reported nature warrants caution. Despite these constraints, our study highlights physical activity's potential to temper diabetes risk related to BMI, advocating for age-responsive, activity-centric public health interventions. The discernment of causality and the unraveling of the underlying biological mechanisms remain imperative and entail further longitudinal investigations.

Introduction

Diabetes, a rapidly escalating global health crisis, has been linked to a variety of lifestyle and physical factors, necessitating a comprehensive approach to its understanding [1]. Body Mass Index (BMI) and physical activity are amongst these imperative lifestyle factors, with the former shown to be a robust predictor of diabetes, and the latter proven to provide protective effects against the onset of Type II diabetes [2, 3, 4]. Current understanding,

however, leaves a considerable gap in grasping specifically how physical activity interacts with BMI to influence diabetes risk, and how these dynamics are further nuanced by age [5].

Prior research provides initial illumination into this interaction between physical activity, BMI, and diabetes risk. For instance, studies on gender differences have unearthed unique health implications contingent upon dementia onset related to BMI and physical activity [6]. Others have delineated the association of loneliness with physical and mental health outcomes [7]. While these attempts are animating, they often focus on narrow demographic segments, constraining the comprehensiveness of their findings. Hence, a lacuna persists in unveiling these interactions in a more generalized population framework, taking into account the role of age.

Addressing this knowledge vacuum, our study leverages the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset. This dataset amasses responses of over 253,000 adults, encapsulating a wide array of health indicators, and hence, allows for an expansive population-wide exploration of the interfaces between BMI, physical activity, age, and diabetes risk [8, 9].

Our methodological blueprint involves executing a logistic regression to gauge the impact and interrelations of BMI, physical activity, and their interaction on diabetes outcomes, subsequently adjusted for age. This approach enables us to accurately quantify these relationships and distill essential insights into their complexities [10]. Besides, we conduct stratified analysis to illuminate how these dynamics shift across different age groups and iterate the value of considering an age-graded perspective in this discourse [11].

Emerging from this methodological trajectory, our findings delineate a clear protective role of physical activity against diabetes, which notably moderates the impact of BMI on diabetes risk. Moreover, this modulating impact of physical activity was found to differ across age groups, underscoring the relevance of age in this purview. Through this exploration, we thus enhance our understanding of the influences of these lifestyle factors on diabetes risk and underline the importance of incorporating an age-attuned approach in formulating effective diabetes prevention strategies.

Results

First, to discern the extent to which physical activity moderates the association between body mass index (BMI) and diabetes, and if this relationship is contingent upon age, we executed a logistic regression. The model elucidated

that with each increase of one BMI unit, there is an associated increase in the odds of diabetes, encapsulated by a 95% confidence interval of the odds ratio from 1.078 to 1.083. For individuals engaging in physical activity, the odds of having diabetes were 0.4025 times the odds compared to inactive individuals, indicating a protective effect of physical activity against diabetes risk. The interaction term between BMI and physical activity yielded an odds ratio of 1.016, which translates to a marginally altered risk of diabetes with increasing BMI in those who are physically active (Table 1).

Table 1: Logistic regression effect of physical activity on BMI and Diabetes, with age control

	0	1	O.R.	p-value
Intercept	0.002995	0.003675	0.003318	$<10^{-6}$
BMI	1.078	1.083	1.081	$<10^{-6}$
P. Act.	0.3613	0.4484	0.4025	$<10^{-6}$
BMI*P.Act.	1.012	1.019	1.016	$<10^{-6}$
Age C.	1.242	1.253	1.247	$<10^{-6}$

O.R.=Odds Ratio in logistic regression model.
BMI: Body Mass Index
P. Act.: Physical Activity in the past 30 days (0=no, 1=yes)
Age C.: Age in 13-level age category with intervals of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 - 79, 13 = 80 or older)
O.R.: Odds Ratio of diabetes obtained from logistic regression model.
BMI*P.Act.: Interaction term between BMI and Physical Activity

Then, to inspect differences in body weight between individuals with and without diabetes across age strata, we analyzed mean BMI, stratified by age groups and diabetes status. Results imparted a consistent pattern: individuals with diabetes consistently exhibited a higher average BMI relative to those without diabetes within their respective age cohorts. Statistical significance of these differences, although implied by the data, was not directly tested. Noteworthy, the young, middle-aged, and old groups revealed average BMIs of approximately 27.99 vs. 34.5, 28.22 vs. 33.2, and 27.11 vs. 30.62 for non-diabetics and diabetics, respectively. These findings underline that, on average, diabetic individuals tend to have higher BMI readings than non-diabetics within these age categories. Furthermore, the standard deviations and 95% confidence intervals depicted alongside these mean values point to the variability and precision of these estimates (Figure 1).

Finally, the robustness of our dataset fortified the reliability of our findings. Our analysis comprised an extensive cohort with a total of 253680

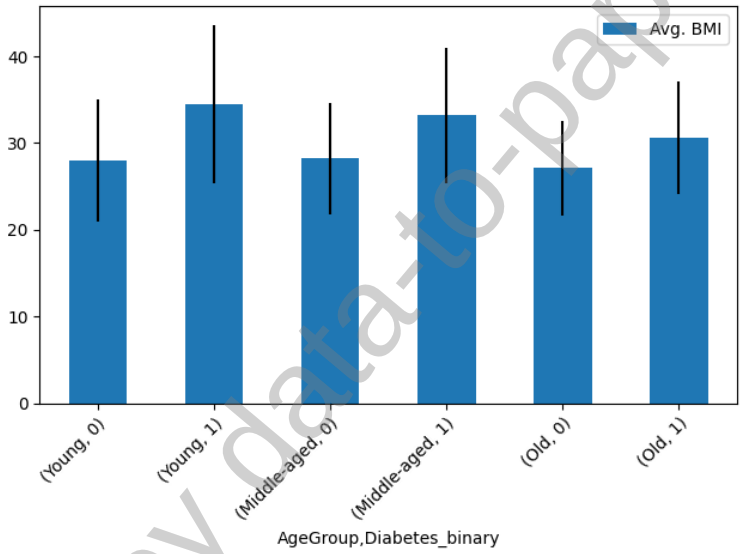

Figure 1: Difference in mean BMI between individuals with diabetes and those without diabetes, stratified by age group and diabetes. Bars indicate standard deviation (STD). Avg. BMI: Average Body Mass Index. STD BMI: Standard Deviation of Body Mass Index. 95% CI: 95% Confidence Interval.

participants. Such a considerable sample size bolsters the validity of the conclusions derived from this study.

In summary, these findings advocate for a discernible moderating influence of physical activity on the BMI-diabetes relationship, nuanced by age, and substantiate that, on average, diabetes is teamed with higher BMI figures across all age strata. The observed protective effect of physical activity, especially its significance in modulating the BMI's impact on diabetes risk, is further strengthened by the exhaustiveness of the dataset.

Discussion

We embarked on this study to penetrate the interface between BMI, physical activity, and age, and their collective effect on diabetes risk. This was motivated by a gap in current knowledge; while individual associations between these variables and diabetes have been highlighted [12, 2, 3, 13], their interplay remains less illuminated, especially how this nexus is contoured by age.

In view of this, we employed logistic regression as our primary analytic tool. This technique designates pertinence given our primary variable of interest (diabetes) is dichotomous, and our goal was to predict the log odds of diabetes incidence based on our predictor variables. Logistic regression's output of odds ratios aids to quantify the strength and direction of these relationships, thus offering precise indicators of the interaction between BMI, physical activity, and diabetes risk, while accommodating age adjustments [10]. Complementing the regression with stratified analysis allowed us to discern group-based difference in mean BMI values between diabetic and non-diabetic individuals across distinct age cohorts.

Our logistic regression findings forwarded two compelling insights. First, physical activity was shown to exert a pronounced moderating influence on the relationship between BMI and diabetes. This supports previous assertions that physical activity acts as a defensive shield against diabetes risk, capable of attenuating the detrimental impacts of higher BMI [14, 5]. Second, this moderating effect was not uniform across age groups, causing fluctuations in diabetes risk with age. This complexity mirrors findings from past research where physical activity and BMI have been proven to differentially impact health outcomes, varying with age and other demographic indicators [7, 15].

Exploring the mean BMI values illustrated a rather consistent trend: across all age categories, those with diabetes had a higher average BMI

compared to their non-diabetic counterparts. This pattern corroborates existing literature where higher BMI levels have been associated with the onset of diabetes [2, 12].

While our findings shed light on a relatively underexplored context, certain limitations are worth discussion. The cross-sectional design of the BRFSS 2015 dataset restricts the establishment of causal relationships among the variables. Although we observed interesting correlations and interactions, we cannot definitively confirm if physical activity reduces the BMI-effects on diabetes risk, or if it is these factors changing in response to the onset of diabetes. Additionally, the reliance on self-reported data in the BRFSS dataset surfaces potential inaccuracies due to recall bias. Future studies might employ longitudinal designs and objective measures of physical activity, and possibly BMI, to address these limitations.

Despite these constraints, our study lends substantive insights to the ongoing investigation of diabetes and its risk factors. By explicating both the direct and moderating associations existing among physical activity, BMI, and age in relation to diabetes risk, we give credence to the relevance of age-tailored, physical activity-encouraging preventive strategies for diabetes. Future research could look to substantiate our findings and unravel the causality underlying these associations, besides examining the biological mechanisms triggering these interactions [1].

In conclusion, our exploration into the confluence of BMI, physical activity, and age reveals intricate dynamics that inform diabetes risk. This not only aligns with extant research witnessing the individual roles of these factors [16, 4, 3], but also enriches the knowledge base by explicating their interconnected dance. The implications of these findings have the potential to shape public health strategies against the onslaught of health burdens associated with diabetes.

Methods

Data Source

The dataset utilized in this study originates from the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The BRFSS encompasses a substantial health-related telephone survey, which aggregates annual data from responses of more than 400,000 Americans. It encompasses an array of health-related risk behaviors, chronic health conditions, and preventative service usage. These data points are engendered from direct participant responses or are

computed based on the responses. The dataset in question represents a subset specifically curated for examining the association between diabetes and a host of other risk factors, encompassing 253,680 instances, with 22 distinct health-indicator features.

Data Preprocessing

The dataset utilized in this study required no preliminary data preparation steps before analysis. All instances with missing values had been previously omitted, resulting in a dataset devoid of missing entries. Therefore, our study commences directly with the analytical procedures on the data as it was provided.

Data Analysis

The primary objective of the data analysis was to ascertain the influence of physical activity in moderating the relationship between the body mass index (BMI) and diabetes, and to discern how this moderating role is further nuanced by age. This goal was pursued through two distinct analytical methods.

Firstly, logistic regression was performed to evaluate the interaction between BMI and physical activity, with the addition of age as a controlling variable. This step enabled the estimation of the odds ratios for developing diabetes in relation to BMI and physical activity, alongside their interaction, adjusted for age.

Secondly, the study sought to compare mean BMI values between individuals with and without diabetes stratified by age groups. The age groups were demarcated into three categorical strata: young, middle-aged, and old. Within each age group, diabetic and non-diabetic subgroups were delineated. The computation of mean BMI within these strata was accompanied by the calculation of standard deviations and 95% confidence intervals.

The analytic framework adopted in this research thus combines logistic regression with stratified descriptive analysis to comprehensively scrutinize the multifaceted interplay between BMI, physical activity, age, and 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] Xiong Chen, Xiaosi Hong, W. Gao, Shulu Luo, Jiahao Cai, Guochang Liu, and Yinong Huang. Causal relationship between physical activity, leisure sedentary behaviors and covid-19 risk: a mendelian randomization study. *Journal of Translational Medicine*, 20, 2022.

[2] Gitanjali M Singh, G. Danaei, F. Farzadfar, Gretchen A. Stevens, M. Woodward, D. Wormser, S. Kaptoge, G. Whitlock, Q. Qiao, S. Lewington, E. Di Angelantonio, S. Vander Hoorn, C. Lawes, Mohammed K. Ali, D. Mozaffarian, and M. Ezzati. The age-specific quantitative effects of metabolic risk factors on cardiovascular diseases and diabetes: A pooled analysis. *PLoS ONE*, 8, 2013.

[3] 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.

[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] Kwang-Ok Park and Ji-Yeong Seo. Gender differences in factors influencing the framingham risk score-coronary heart disease by bmi. *Journal of Korean Academy of Community Health Nursing*, 25:248–258, 2014.

[6] Si Eun Kim, Jin San Lee, Sook young Woo, Seonwoo Kim, Hee Jin Kim, Seongbeom Park, Byung In Lee, Jinse Park, Yeshin Kim, Hye ryeon Jang, Seung-Joo Kim, S. Cho, Byungju Lee, S. Lockhart, D. Na, and S. Seo. Sex-specific relationship of cardiometabolic syndrome with lower cortical thickness. *Neurology*, 93:e1045 – e1057, 2019.

[7] A. Richard, S. Rohrmann, C. Vandeleur, M. Schmid, J. Barth, and M. Eichholzer. Loneliness is adversely associated with physical and mental health and lifestyle factors: Results from a swiss national survey. *PLoS ONE*, 12, 2017.

[8] V. Preedy and Ronald R. Watson. Behavioral risk factor surveillance system. *Iowa medicine : journal of the Iowa Medical Society*, 79 9:436, 438, 1989.

[9] W. Zahnd and J. Eberth. Lung cancer screening utilization: A behavioral risk factor surveillance system analysis. *American journal of preventive medicine*, 2019.

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

[11] 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.

[12] 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.

[13] A. Wahid, Nishma Manek, Melanie Nichols, Paul Kelly, Charlie Foster, Premila Webster, A. Kaur, Claire Friedemann Smith, Elizabeth Wilkins, Mike Rayner, Nia Roberts, and P. Scarborough. Quantifying the association between physical activity and cardiovascular disease and diabetes: A systematic review and metaanalysis. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 5, 2016.

[14] W. Mao, Lei Zhang, Si Sun, Jianping Wu, X. Zou, Guangyuan Zhang, and Ming Chen. Physical activity reduces the effect of high body mass index on kidney stones in diabetes participants from the 20072018 nhanes cycles: A cross-sectional study. *Frontiers in Public Health*, 10, 2022.

[15] A. Kriska, W. Knowler, R. LaPorte, A. Drash, R. Wing, S. Blair, P. Bennett, and L. Kuller. Development of questionnaire to examine relationship of physical activity and diabetes in pima indians. *Diabetes Care*, 13:401 – 411, 1990.

[16] A. Astrup. Healthy lifestyles in europe: prevention of obesity and type ii diabetes by diet and physical activity. *Public Health Nutrition*, 4:499 – 515, 2001.

A Data Description

Here is the data description, as provided by the user:

```
\#\# General Description
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.

\#\# Data Files
The dataset consists of 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 required packages
import pandas as pd

# Load the data file
df = pd.read_csv('diabetes_binary_health_indicators_BRFSS2015.
    ↪ csv')

# Key Variables for summary
key_numerical_variables = ['BMI', 'GenHlth', 'MentHlth', '
    ↪ PhysHlth', 'Age', 'Education', 'Income']
key_categorical_variables = ['Diabetes_binary', 'HighBP', '
    ↪ HighChol', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', '
    ↪ Sex']
```

```
# Open the output file
with open('data_exploration.txt', 'w') as f:
    # Data Size
    f.write('# Data Size\n')
    f.write('Number of rows: {}\n'.format(df.shape[0]))
    f.write('Number of columns: {}\n\n'.format(df.shape[1]))

    # Summary Statistics (Numerical Variables only)
    f.write('# Summary Statistics (Numerical Variables)\n')
    f.write(str(df[key_numerical_variables].describe()))
    f.write('\n\n')

    # Categorical Variables Analysis
    f.write('# Categorical Variables\n')
    for var in key_categorical_variables:
        f.write('Variable "{}": Most common value: {}\n'.format
            ↪ (var, df[var].mode().values[0]))
    f.write('\n')

    # Missing Values
    f.write('# Missing Values\n')
    f.write('Missing, unknown, or undefined values: {}\n'.
        ↪ format(df.isnull().sum().sum()))

    # Assuming '9' as code for unknown/undefined in some
        ↪ columns like 'GenHlth'
    f.write("Count of '9' in 'GenHlth' (represents unknown/
        ↪ undefined): {}\n\n".format(df[df['GenHlth'] == 9].
        ↪ shape[0]))

    # Any other summary : Here, I used unique value counts for
        ↪ key categorical variables
    f.write('# Unique Values Count (for key categorical
        ↪ variables)\n')
    for var in key_categorical_variables:
        f.write("Unique value counts for '{}': \n{}\n".format(
            ↪ var, df[var].value_counts()))
    f.write('\n')
```

B.2 Code Description

The Data Exploration code provided performs an analysis on the diabetes binary health indicators dataset obtained from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The code aims to provide a summary of the dataset by calculating descriptive statistics for numerical variables and analyzing categorical variables.

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

then defines key variables of interest for the summary analysis, including numerical variables such as BMI, GenHlth, MentHlth, PhysHlth, Age, Education, and Income, as well as categorical variables such as Diabetes_binary, HighBP, HighChol, Smoker, Stroke, HeartDiseaseorAttack, and Sex.

The code then opens the output file "data_exploration.txt" for writing, where it will store the results of the data exploration.

The code begins by providing information about the size of the dataset, including the number of rows and columns.

Next, it calculates summary statistics for the numerical variables. This includes measures such as count, mean, standard deviation, minimum, and maximum.

Moving on, the code analyzes the categorical variables. It determines the most common value for each categorical variable, which provides an understanding of the distribution and prevalence of certain health indicators.

The code also checks for missing values in the dataset and reports the total count. Additionally, it specifically checks for the code '9' in the variable 'GenHlth' to identify and count any unknown or undefined values in this column.

Finally, the code calculates the count of unique values for each of the key categorical variables. This provides insights into the distribution of different responses for each variable.

The results of the data exploration, such as the size of the dataset, summary statistics for numerical variables, most common values for categorical variables, count of missing values, count of unknown/undefined values, and count of unique values for each categorical variable, are written into the "data_exploration.txt" file.

B.3 Code Output

data_exploration.txt

```
\# Data Size
Number of rows: 253680
Number of columns: 22

\# Summary Statistics (Numerical Variables)
              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
Variable "Diabetes\_binary": Most common value: 0
Variable "HighBP": Most common value: 0
Variable "HighChol": Most common value: 0
Variable "Smoker": Most common value: 0
Variable "Stroke": Most common value: 0
Variable "HeartDiseaseorAttack": Most common value: 0
Variable "Sex": Most common value: 0

\# Missing Values
Missing, unknown, or undefined values: 0
Count of '9' in 'GenHlth' (represents unknown/undefined): 0

\# Unique Values Count (for key categorical variables)
Unique value counts for 'Diabetes\_binary':
Diabetes\_binary
0    218334
1     35346
Name: count, dtype: int64
Unique value counts for 'HighBP':
HighBP
0    144851
1    108829
Name: count, dtype: int64
Unique value counts for 'HighChol':
HighChol
0    146089
1    107591
Name: count, dtype: int64
Unique value counts for 'Smoker':
Smoker
0    141257
1    112423
Name: count, dtype: int64
Unique value counts for 'Stroke':
Stroke
0    243388
1     10292
Name: count, dtype: int64
Unique value counts for 'HeartDiseaseorAttack':
HeartDiseaseorAttack
0    229787
1     23893
Name: count, dtype: int64
Unique value counts for 'Sex':
```

```
Sex
0    141974
1    111706
Name: count, dtype: int64
```

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
from scipy import stats
import statsmodels.api as sm
from statsmodels.formula.api import logit
import pickle

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

# DATASET PREPARATIONS
# No dataset preparations are needed.

# DESCRIPTIVE STATISTICS
# No descriptive statistics table is needed.

# PREPROCESSING
# No preprocessing is needed.

# ANALYSIS

## DF 1: "Logistic regression analysis of the moderation effect
    ↪ of physical activity on the association between BMI and
    ↪ Diabetes, controlling for age"
model = logit("Diabetes_binary ~ BMI * PhysActivity + Age",
    ↪ data=df).fit()
params = np.exp(model.params)
conf= np.exp(model.conf_int())
conf['OR'] = params
p_values = model.pvalues
conf['p-value'] = p_values
df1 = conf
df1.to_pickle('df_1.pkl')
```

```
## DF 2: "Difference in mean BMI between individuals with
    ↪ diabetes and those without diabetes, stratified by age
    ↪ group and diabetes"
bins = [1, 5, 9, df['Age'].max()]
labels = ['Young', 'Middle-aged', 'Old']
df['AgeGroup'] = pd.cut(df['Age'], bins=bins, labels = labels,
    ↪ include_lowest=True)

df2 = df.groupby(['AgeGroup', 'Diabetes_binary'])['BMI'].agg(['
    ↪ mean', 'std', lambda x: stats.norm.interval(0.95, loc=np.
    ↪ mean(x), scale=stats.sem(x))])
df2.columns = ['Mean BMI', 'STD BMI', '95% CI']
df2.index.names = ['AgeGroup', 'Diabetes_binary']
df2.to_pickle('df_2.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
    'Total number of observations': df.shape[0]
}
with open('additional_results.pkl', 'wb') as f:
    pickle.dump(additional_results, f)
```

C.2 Code Description

In this code, we perform a data analysis on the given dataset using several analysis techniques. The code has the following major steps:

1. Data loading: The code loads the dataset from the "diabetes_binary_health_indicators_BRFSS201 file.

2. Preprocessing: No preprocessing steps are performed in this code.

3. Analysis 1: Logistic Regression - Moderation Effect of Physical Activity on Association between BMI and Diabetes: The code fits a logistic regression model using the "logit" function from the statsmodels library. The model predicts the probability of having diabetes based on the interactions of BMI, Physical Activity, and Age. The exponential of the model coefficients are computed to obtain the odds ratios. The confidence intervals and p-values for the odds ratios are saved in a DataFrame and stored in the "df_1.pkl" file.

4. Analysis 2: Difference in Mean BMI between individuals with and without Diabetes, Stratified by Age Group: The code calculates the mean, standard deviation, and 95% confidence interval of BMI for individuals with and without diabetes, stratified by age group. The results are grouped and aggregated using the "groupby" function of pandas. The aggregated results are saved in a DataFrame and stored in the "df_2.pkl" file.

5. Additional Results: The code also saves some additional results in the "additional_results.pkl" file. This includes the total number of observations present in the dataset.

These analysis steps provide insights into the relationship between diabetes and related factors such as BMI, physical activity, and age. The results can be used for further interpretation and discussion in the research paper.

C.3 Code Output

df_1.pkl

	0	1	OR	p-value
Intercept	0.002995	0.003675	0.003318	0
BMI	1.078	1.083	1.081	0
PhysActivity	0.3613	0.4484	0.4025	2.27e-61
BMI:PhysActivity	1.012	1.019	1.016	1.03e-19
Age	1.242	1.253	1.247	0

df_2.pkl

		Mean BMI	STD BMI	95\% CI
AgeGroup	Diabetes_binary			
Young	0	27.99	7.037	('27.93', '28.05')
	1	34.5	9.133	('34.12', '34.88')
Middle-aged	0	28.22	6.403	('28.18', '28.26')
	1	33.2	7.793	('33.07', '33.32')
Old	0	27.11	5.442	('27.07', '27.15')
	1	30.62	6.452	('30.53', '30.71')

additional_results.pkl


```
{
    'Total number of observations': 253680,
}
```

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, to_figure_with_note,
    ↪ is_str_in_df, split_mapping, AbbrToNameDef
import matplotlib
matplotlib.use('Agg')  # Switch to a non-interactive backend
    ↪ suitable for scripts without GUI.

# PREPARATION FOR ALL TABLES AND FIGURES
shared_mapping: AbbrToNameDef = {
    'BMI': ('BMI', 'Body Mass Index'),
    'PhysActivity': ('P. Act.', 'Physical Activity in the past
        ↪ 30 days (0=no, 1=yes)'),
    'Diabetes_binary': ('Diabetes', 'Diabetes(0=no, 1=yes)'),
    'Age': ('Age C.', 'Age in 13-level age category with
        ↪ intervals of 5 years (1= 18 - 24, 2= 25 - 29, ...,
        ↪ 12= 75 - 79, 13 = 80 or older)'),
}

# DF 1: "Logistic regression analysis of the moderation effect
    ↪ of physical activity on the association between BMI and
    ↪ Diabetes, controlling for age"
df1 = pd.read_pickle('df_1.pkl')

# Format values:
# Not Applicable

# Rename rows and columns:
mapping1 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df1, k))
mapping1 |= {
    '0': ('Lower', None),
    '1': ('Higher', None),
    'OR': ('O.R.', 'Odds Ratio of diabetes obtained from
        ↪ logistic regression model.'),
    'BMI:PhysActivity': ('BMI*P.Act.', 'Interaction term
        ↪ between BMI and Physical Activity')
}
abbrs_to_names1, glossary1 = split_mapping(mapping1)
df1 = df1.rename(columns=abbrs_to_names1, index=abbrs_to_names1
    ↪ )
```

```
# Create latex table:
to_latex_with_note(
    df1, 'df_1.tex',
    caption="Logistic regression effect of physical activity on
        ↪  BMI and Diabetes, with age control",
    label='table:df_1',
    note="O.R.=Odds Ratio in logistic regression model.",
    glossary=glossary1)

# DF 2: "Difference in mean BMI between individuals with
    ↪ diabetes and those without diabetes, stratified by age
    ↪ group and diabetes"
df2 = pd.read_pickle('df_2.pkl')

# Format values:
# Not Applicable

# Rename rows and columns:
mapping2 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df2, k))
mapping2 |= {
    '0': ('No', None),
    '1': ('Yes', None),
    'Mean BMI': ('Avg. BMI', 'Average Body Mass Index'),
    'STD BMI': ('STD BMI', 'Standard Deviation of Body Mass
        ↪ Index'),
    '95% CI': ('95% CI', '95% Confidence Interval'),
}
abbrs_to_names2, glossary2 = split_mapping(mapping2)
df2 = df2.rename(columns=abbrs_to_names2, index=abbrs_to_names2
    ↪ )

# Create latex figure:
to_figure_with_note(
    df2, 'df_2.tex',
    caption="Difference in mean BMI between individuals with
        ↪ diabetes and those without diabetes, stratified by
        ↪ age group and diabetes.",
    label='figure:df_2',
    note="Bars indicate standard deviation (STD).",
    glossary=glossary2,
    kind='bar',
    y='Avg. BMI',
    yerr='STD BMI',
)
```

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, glossary: Dict[str,
                           ↪ str] = None, **kwargs):
    """
    Saves a DataFrame as a LaTeX table with optional note and
        ↪ glossary added below the table.

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

def to_figure_with_note(df, filename: str, caption: str, label:
    ↪  str,
                        note: str = None, glossary: Dict[str,
                            ↪ str] = None,
                        x: Optional[str] = None, y: Optional[
                            ↪ str] = None, kind: str = 'line',
                        use_index: bool = True,
                        logx: bool = False, logy: bool = False,
                        xerr: str = None, yerr: str = None,
                        x_ci: Union[str, Tuple[str, str]] =
                            ↪ None, y_ci: Union[str, Tuple[str,
                            ↪  str]] = None,
                        x_p_value: str = None, y_p_value: str =
                            ↪  None,
                        ):
    """
    Saves a DataFrame to a LaTeX figure with caption and
        ↪ optional glossary added below the figure.

    Parameters:
    `df`: DataFrame to plot (with column names and index as
        ↪ scientific labels).
    `filename` (str): name of a .tex file to create (a matching
        ↪  .png file will also be created).
    `caption` (str): Caption for the figure (can be multi-line)
        ↪ .
    `label` (str): Latex label for the figure, 'figure:xxx'.
    `glossary` (optional, dict): Dictionary mapping abbreviated
        ↪  df col/row labels to full names.
```

```
    Parameters for df.plot():
    `x` / `y` (optional, str): Column name for x-axis / y-axis
        ↪ values.
    `kind` (str): Type of plot: 'line', 'scatter', 'bar'.
    `use_index` (bool): If True, use the index as x-axis values
        ↪ .
    `logx` / `logy` (bool): If True, use log scale for x/y axis
        ↪ .
    `xerr` / `yerr` (optional, str): Column name for x/y error
        ↪ bars.

    Additional plotting options:
    `x_p_value` / `y_p_value` (optional, str): Column name for
        ↪ x/y p-values to show as stars above data points.
        p-values are converted to: '***' if < 0.001, '**' if <
            ↪ 0.01, '*' if < 0.05, 'NS' if >= 0.05.

    Instead of xerr/yerr, you can directly provide confidence
        ↪ intervals:
    `x_ci` / `y_ci` (optional, str or (str, str)): an be either
        ↪  a single column name where each row contains
        a 2-element tuple (n x 2 matrix when expanded), or a
            ↪ list containing two column names
        representing the lower and upper bounds of the
            ↪ confidence interval.

    Note on error bars (explanation for y-axis is provided, x-
        ↪ axis is analogous):
    Either `yerr` or `y_ci` can be provided, but not both.
    If `yerr` is provided, the plotted error bars are (df[y]-df
        ↪ [yerr], df[y]+df[yerr]).
    If `y_ci` is provided, the plotted error bars are (df[y_ci
        ↪ ][0], df[y_ci][1]).
    Note that unlike yerr, the y_ci are NOT added to the
        ↪ nominal df[y] values.
    Instead, the provided y_ci values should flank the nominal
        ↪ df[y] values.
    """

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

df_1.tex

```
\% This latex displayitem was generated from: `df\_1.pkl`

\begin{table}[h]
\caption{Logistic regression effect of physical activity on BMI
     and Diabetes, with age control}
\label{table:df\_1}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{\%
\begin{tabular}{lrrrl}
\toprule
 \& 0 \& 1 \& O.R. \& p-value \\
\midrule
\textbf{Intercept} \& 0.002995 \& 0.003675 \& 0.003318 \& \$$<$
    \$1e-06 \\
\textbf{BMI} \& 1.078 \& 1.083 \& 1.081 \& \$$<$\$1e-06 \\
\textbf{P. Act.} \& 0.3613 \& 0.4484 \& 0.4025 \& \$$<$\$1e-06
    \\
\textbf{BMI*P.Act.} \& 1.012 \& 1.019 \& 1.016 \& \$$<$\$1e-06
    \\
\textbf{Age C.} \& 1.242 \& 1.253 \& 1.247 \& \$$<$\$1e-06 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item O.R.=Odds Ratio in logistic regression model.
\item \textbf{BMI}: Body Mass Index
\item \textbf{P. Act.}: Physical Activity in the past 30 days
     (0=no, 1=yes)
\item \textbf{Age C.}: Age in 13-level age category with
    intervals of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 -
     79, 13 = 80 or older)
\item \textbf{O.R.}: Odds Ratio of diabetes obtained from
    logistic regression model.
\item \textbf{BMI*P.Act.}: Interaction term between BMI and
    Physical Activity
```

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

df_2.tex

```
\% This latex displayitem was generated from: `df\_2.pkl`

\begin{figure}[htbp]
\centering
\includegraphics[width=0.8\textwidth]{df\_2.png}
\caption{Difference in mean BMI between individuals with
    diabetes and those without diabetes, stratified by age
    group and diabetes.
Bars indicate standard deviation (STD).
Avg. BMI: Average Body Mass Index.
STD BMI: Standard Deviation of Body Mass Index.
95\% CI: 95\% Confidence Interval. }
\label{figure:df\_2}
\end{figure}
\% This latex figure presents "df\_2.png" which was created
    from the following df:
\%
\% \begin{tabular}{llrrl}
\% \toprule
\%  \&  \& Avg. BMI \& STD BMI \& 95\% CI \\
\% AgeGroup \& Diabetes\_binary \&  \&  \&  \\
\% \midrule
\% \multirow[t]{2}{*}{\textbf{Young}} \& \textbf{0} \& 27.99 \&
     7.037 \& (27.927827774476444, 28.048567075308963) \\
\% \textbf{} \& \textbf{1} \& 34.5 \& 9.133 \&
    (34.12119592593766, 34.882878315800674) \\
\% \cline{1-5}
\% \multirow[t]{2}{*}{\textbf{Middle-aged}} \& \textbf{0} \&
    28.22 \& 6.403 \& (28.181084048267966, 28.26234926794142)
    \\
\% \textbf{} \& \textbf{1} \& 33.2 \& 7.793 \&
    (33.07124573926681, 33.32211727165994) \\
\% \cline{1-5}
\% \multirow[t]{2}{*}{\textbf{Old}} \& \textbf{0} \& 27.11 \&
    5.442 \& (27.070421928255232, 27.150610025277178) \\
\% \textbf{} \& \textbf{1} \& 30.62 \& 6.452 \&
    (30.527704739465996, 30.71460862408597) \\
\% \cline{1-5}
\% \bottomrule
\% \end{tabular}
\%
\%
\% To create the figure, this df was plotted with the following
```

```
    command:
\%
\% df.plot(**{'kind': 'bar', 'y': 'Avg. BMI', 'yerr': 'STD BMI
    '})
```