Physical Activity, BMI, and Age: Impacts on Diabetes Risk in a National Study

Data to Paper

September 28, 2023

Abstract

Diabetes is a prevalent chronic health condition with significant societal and individual burdens. Understanding the factors influencing the development of diabetes is crucial for effective prevention and management. However, a crucial research gap exists regarding the complex interplay between physical activity, body mass index (BMI), age, and their interaction effects on diabetes risk. In this study, we analyzed data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015, comprising over 250,000 individuals in the United States, to address this research gap. By employing logistic regression analysis and adjusting for potential confounders, we found that lower physical activity levels, higher BMI, and older age were associated with an increased likelihood of developing diabetes. Furthermore, the results revealed significant interaction effects between physical activity and BMI, as well as physical activity and age, suggesting potential synergistic impacts on diabetes risk. While this study is observational in nature, limiting our ability to establish causality, the findings offer valuable insights into the associations between physical activity, BMI, age, and the development of diabetes. These results highlight the importance of promoting physical activity, maintaining a healthy BMI, and encouraging healthy lifestyle choices across different age groups to potentially reduce the burden of diabetes and improve population health.

Introduction

Diabetes is a chronic disease with widespread impact that poses significant societal and individual challenges [1, 2]. Considerable strides have been made in understanding the factors influencing diabetes development, providing a foundation for targeted prevention and management efforts. Phys-

ical activity, body mass index (BMI), and age, among other factors, have been widely documented as influential factors in the development of diabetes [3, 4].

Existing literature, while rich, has often considered these factors in isolation, resulting in established individual relationships [5]. For instance, a strong evidence base exists indicating an inverse association between physical activity and diabetes risk [6]. Also, research has demonstrated a positive association between BMI and diabetes risk, especially among younger adults [7]. However, less is known regarding the complexities of the interaction among physical activity, BMI, and age and their collective effect on diabetes risk.

The present study addresses this research gap by exploring the synergistic effects of physical activity, BMI, and age on diabetes risk, using a comprehensive data set of the Behavioral Risk Factor Surveillance System (BRFSS) from the CDC for the year 2015 [8]. The dataset, extracted from responses of over 400,000 American participants, provides an opportunity for population-level investigations enhancing the understanding of the interconnected dynamics of the three factors in the context of diabetes risk.

Employing logistic regression analysis, we investigated the associations between physical activity, BMI, and age with diabetes risk, adjusting for potential confounders such as sex, education level, and income [9, 10]. Our analysis reveals significant interactions between physical activity and BMI, as well as physical activity and age on diabetes risk. The findings provide valuable insights into the complexities of diabetes risk factors, highlighting the need to consider the synergistic impacts of physical activity, BMI, and age in the ongoing efforts to curb the rising prevalence of diabetes.

Results

To investigate the relationship between physical activity, BMI, age, and diabetes risk, we conducted a comprehensive analysis using data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015, comprising 253,626 observations.

Firstly, to understand the general patterns of diabetes prevalence, we examined the mean distribution of physical activity levels, BMI, and Age, among individuals with and without diabetes (Table 1). The results revealed that, individuals with diabetes tend to have less physical activity (0.631) and have higher BMI (31.9), compared to the individuals without diabetes whose mean physical activity was 0.777 and BMI was 27.8. Additionally,

the average age was higher among individuals diagnosed with diabetes (9.38) compared to those without (7.81).

Table 1: Mean values of physical activity, BMI, age, and potential confounders stratified by diabetes status

	Physical Activity	BMI	Age	Sex	Education Level	Income level
No Diabetes	0.777	27.8	7.81	0.434	5.1	6.19
Diabetes	0.631	31.9	9.38	0.479	4.75	5.21

Physical Activity: Physical Activity in past 30 days, 1: Yes, 0: No
BMI: Body Mass Index, kg/m2
Age: 13-level age category in 5 years intervals
Sex: 0: Female, 1: Male
Education Level: 1: Never attended school, 2: Elementary, 3: Some high school, 4: High school, 5: Some college, 6: College
Income level: Income level on a scale of 1-8 (1: <=10K, 2: <=15K, 3: <=20K, 4: <=25K, 5: <=35K, 6: <=50K, 7: <=75K, 8: >75K)

Secondly, to test the effect of physical activity, BMI, and age on the risk of diabetes, we built a multiple logistic regression model adjusting for sex, education level and income (Table 2). The results showed that each unit decrease in physical activity is associated with increased log-odds of having diabetes by -0.97, and similarly for each unit increase in BMI and Age, the log-odds increased by 0.0768 and 0.188 respectively.

Finally, to further examine the interactive effect of physical activity with BMI and Age, we included interaction terms in our regression model (Table 2). The coefficient for the interaction term between physical activity and BMI is 0.0137 and between physical activity and age is 0.027, both significant with p-value $< 10^{-6}$. This suggests that lower levels of physical activity combined with higher BMI or age could have a compounded effect on the risk of diabetes.

Taken together, these results highlight the significant effects of physical activity, BMI, and age on the risk of diabetes. Notably, the interaction of physical activity with BMI and age suggests that combined influence could be particularly important for the risk of diabetes.

Discussion

In this study, we investigated the interplay between physical activity, BMI, age, and the risk of diabetes [1, 2]. Using a comprehensive dataset derived

Table 2: Logistic regression of diabetes status on physical activity, BMI, age, and their interaction terms, adjusting for sex, education, and income

	coef	std err	pvalue	[0.025	0.975]
Constant	-4.35	0.0716	$<10^{-6}$	-4.49	-4.21
Physical Activity	-0.97	0.0818	$<10^{-6}$	-1.13	-0.81
BMI	0.0768	0.0014	$<10^{-6}$	0.074	0.0795
Age	0.188	0.00411	$<10^{-6}$	0.18	0.196
Sex	0.349	0.0125	$<10^{-6}$	0.325	0.374
Education Level	-0.0997	0.00657	$<10^{-6}$	-0.113	-0.0868
Income level	-0.142	0.00317	$<10^{-6}$	-0.148	-0.136
Physical Activity * BMI	0.0137	0.0018	$<10^{-6}$	0.0102	0.0173
Physical Activity * Age	0.027	0.00501	$<10^{-6}$	0.0172	0.0368

pvalue: If value is less than 10^{-6}, it is represented as less than 10^{-6}
Physical Activity: Physical Activity in past 30 days, 1: Yes, 0: No
BMI: Body Mass Index, kg/m2
Age: 13-level age category in 5 years intervals
Sex: 0: Female, 1: Male
Education Level: 1: Never attended school, 2: Elementary, 3: Some high school, 4: High school, 5: Some college, 6: College
Income level: Income level on a scale of 1-8 (1: <=10K, 2: <=15K, 3: <=20K, 4: <=25K, 5: <=35K, 6: <=50K, 7: <=75K, 8: >75K)
Physical Activity * BMI: Interaction term between Physical Activity and BMI
Physical Activity * Age: Interaction term between Physical Activity and Age

from the CDC's BRFSS survey for the year 2015, our methodological approach involved logistic regression, which opened pathways to discern the relationship between the targeted variables and the risk of diabetes, factoring in potential confounding variables such as sex, education level, and income [5].

Our findings corroborate extant literature that points to an inverse relationship between physical activity and diabetes risk [6]. Notably, our study augments this knowledge base by revealing that the interaction between BMI, age, and physical activity significantly impacts the risk of diabetes. These findings are consistent with the work of Singh et al. (2013), where the decline in different risk factors with age was substantiated and attributed to the reduction in physical activity common among older adults [5]. Moreover, comparable to the findings of Zhang et al. (2010), we also demonstrate that maintaining healthy BMI and regular physical activity may reduce the risks of diabetes [11].

However, like all research, our study had specific limitations. On a

methodological aspect, the cross-sectional approach used in this study hampers the ability to confirm causality of the observed associations. Future research could employ longitudinal studies, as in the study by Chen et al., to address this causality aspect and reinforce the findings of this current research [7]. Additionally, the reliance on self-reported data via the BRFSS may introduce recall biases, thus affecting the accuracy of the dataset. Future research may address this by adopting data collection methods that could inherently minimize such bias effects. Lastly, this study did not investigate more complex interactions involving other socio-demographic and lifestyle variables which might interact with physical activity, BMI, and age, and influence diabetes risk. Future work could further examine these interactions to yield a more comprehensive understanding of these influential factors.

In conclusion, our study revealed valuable insights into the interaction of physical activity, BMI, and age in relation to the risk of diabetes. These findings reinforce the necessity of health promotion programs that not only encourage regular physical activity and maintenance of healthy BMI across different age groups but also adapt to the specific needs and challenges that each age group might face. As such, these programs could be instrumental in reducing the burden of diabetes and improving population health. The implications transcend beyond the American population from which the dataset was derived, as diabetes is a global health challenge. These insights, therefore, could be inferentially applicable to global health promotion efforts. Further research, particularly from a longitudinal perspective, could substantiate these findings and validate the causative relationships between these factors and diabetes risk.

Methods

Data Source

We obtained the dataset for our study from the Behavioral Risk Factor Surveillance System (BRFSS) conducted by the Centers for Disease Control and Prevention (CDC). The dataset represents a subset of the BRFSS data collected in the year 2015. This annual health-related telephone survey collects responses from over 400,000 Americans on various health-related risk behaviors, chronic health conditions, and the use of preventative services. The dataset includes information on diabetes-related factors, such as diabetes status, physical activity, body mass index (BMI), and age.

Data Preprocessing

We performed data preprocessing to ensure the quality and suitability of the dataset for analysis. We excluded rows with any values over 90, as these were considered undefined or missing data. This step was implemented to ensure the integrity of the dataset and eliminate potential outliers. The resulting dataset, after data cleaning, consisted of 253,680 responses and 22 features.

Data Analysis

To investigate the association between physical activity and the presence of diabetes, as well as the potential moderating effects of BMI and age, we conducted logistic regression analysis. The analysis involved the estimation of a logistic regression model with diabetes status as the dependent variable and physical activity, BMI, age, and their interaction terms as independent variables.

We calculated mean values of physical activity, BMI, age, and potential confounders stratified by diabetes status to provide descriptive statistics of the study population. The statistical significance of differences between groups was assessed using appropriate statistical tests.

For the logistic regression analysis, we used the cleaned dataset and adjusted for potential confounding factors, including sex, education, and income. We created interaction terms between physical activity and BMI, as well as physical activity and age, to explore potential synergistic effects.

We performed logistic regression using the Statsmodels library in Python. The results of the regression analysis yielded coefficient estimates, standard errors, p-values, and confidence intervals for the predictors. We also calculated the pseudo R-squared of the regression model to assess the proportion of variation explained by the predictors.

All statistical analyses were conducted using Python 3. We used the Pandas library to load and manipulate the dataset, and the Statsmodels library to perform the logistic regression analysis.

Code Availability

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

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 1-8 (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 data
df =
↪  pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.csv")

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

    # Data Size
    f.write("# Data Size\n")
    f.write(f"Number of rows: {df.shape[0]}\n")
```

```
14      f.write(f"Number of columns: {df.shape[1]}\n\n")
15
16      # Summary Statistics
17      f.write("# Summary Statistics\n")
18      desc = df.describe()
19      f.write(f"{desc}\n\n")
20
21      # Categorical Variables
22      f.write("# Categorical Variables\n")
23      categorical_columns =
        ↪  df.select_dtypes(include=['object']).columns
24      for cat_col in categorical_columns:
25          f.write(f"{cat_col} - most common values:\n")
26          f.write(f"{df[cat_col].value_counts().head()}\n\n")
27
28      # Missing Values
29      f.write("# Missing Values\n")
30      missing = df.isnull().sum()
31      f.write(f"{missing}\n")
32
33      # Look for special numeric values that stand for
        ↪  unknown/undefined
34      special_values = df.isin([99, 88, 77]).sum()  #you may
        ↪  adjust these values as needed based on your data set
35      if special_values.sum() > 0:
36          f.write("\nCounts of special numeric values (99,88,77)
            ↪  that stand for unknown/undefined:\n")
37          f.write(f"{special_values}\n")
38
```

B.2 Code Description

The code performs an initial exploration and analysis of the dataset. It starts by loading the dataset into a pandas DataFrame. The code then proceeds to analyze the dataset using several steps.

First, it determines the size of the dataset by writing the number of rows and columns to the "data_exploration.txt" file.

Next, summary statistics are generated using the describe() method on the DataFrame. These statistics include count, mean, standard deviation, minimum, quartiles, and maximum values for each numeric column in the

dataset. The summary statistics are appended to the "data_exploration.txt" file.

Categorical variables are then examined by identifying columns with object data type, indicating categorical variables. For each categorical column, the code writes the most common values and their corresponding counts to the "data_exploration.txt" file.

The code also checks for missing values in the dataset using the isnull() method, which returns a boolean DataFrame indicating whether each value in the dataset is missing or not. The sum() method is then used to calculate the total number of missing values for each column. The results are written to the "data_exploration.txt" file.

Finally, the code checks for special numeric values that stand for unknown or undefined entries. In this code, the values 99, 88, and 77 are considered as special values. The isin() method is used to create a boolean DataFrame where entries equal to these values are marked as True. The sum() method is then used to calculate the total count of each special value for each column. If any special values are found, the counts are written to the "data_exploration.txt" file.

Overall, the code provides a preliminary analysis of the dataset by examining its size, generating summary statistics, identifying common categorical values, detecting missing values, and checking for special numeric values that may affect the data quality.

B.3 Code Output

data_exploration.txt

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

# Summary Statistics
       Diabetes_binary  HighBP  HighChol  CholCheck    BMI  Smoker  Stroke
    HeartDiseaseorAttack  PhysActivity  Fruits  Veggies  HvyAlcoholConsump
    AnyHealthcare  NoDocbcCost  GenHlth  MentHlth  PhysHlth  DiffWalk    Sex    Age
    Education  Income
count           253680  253680    253680     253680 253680  253680  253680
    253680        253680  253680   253680             253680         253680
    253680   253680    253680    253680    253680 253680 253680     253680  253680
mean            0.1393   0.429    0.4241     0.9627  28.38  0.4432 0.04057
```

```
    0.09419        0.7565  0.6343   0.8114             0.0562         0.9511
    0.08418    2.511     3.185     4.242    0.1682 0.4403  8.032       5.05   6.054
std             0.3463  0.4949    0.4942     0.1896  6.609  0.4968  0.1973
    0.2921        0.4292  0.4816   0.3912             0.2303         0.2158
    0.2777    1.068     7.413     8.718    0.3741 0.4964  3.054     0.9858   2.071
min                   0       0         0          0     12       0       0
    0             0       0        0                  0             0            0
    1         0         0         0      0      1          1       1
25%                   0       0         0          1     24       0       0
    0             1       0        1                  0             1            0
    2         0         0         0      0      6          4       5
50%                   0       0         0          1     27       0       0
    0             1       1        1                  0             1            0
    2         0         0         0      0      8          5       7
75%                   0       1         1          1     31       1       0
    0             1       1        1                  0             1            0
    3         2         3         0      1     10          6       8
max                   1       1         1          1     98       1       1
    1             1       1        1                  1             1            1
    5        30        30         1      1     13          6       8

# Categorical Variables
# Missing Values
Diabetes_binary         0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
```

```
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
dtype: int64

Counts of special numeric values (99,88,77) that stand for unknown/undefined:
Diabetes_binary          0
HighBP                   0
HighChol                 0
CholCheck                0
BMI                     57
Smoker                   0
Stroke                   0
HeartDiseaseorAttack     0
PhysActivity             0
Fruits                   0
Veggies                  0
HvyAlcoholConsump        0
AnyHealthcare            0
NoDocbcCost              0
GenHlth                  0
MentHlth                 0
PhysHlth                 0
DiffWalk                 0
Sex                      0
Age                      0
Education                0
Income                   0
dtype: int64
```

C Data Analysis

C.1 Code

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

```python

# IMPORT
import pandas as pd
import pickle
import statsmodels.api as sm

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

# DATASET PREPARATIONS
# Any value over 90 in any of the columns considered to be undefined/missing.
# We will exclude these rows from our analysis.
df_cleaned = df[df.lt(90).all(axis=1)]

# DESCRIPTIVE STATISTICS
## Table 0: "Mean values of physical activity, BMI, age, and potential confounders stratified by diabetes status"
df0 = df_cleaned.groupby('Diabetes_binary')[['PhysActivity', 'BMI', 'Age', 'Sex', 'Education', 'Income']].mean()
df0.index = ['No Diabetes', 'Diabetes']
df0.to_pickle('table_0.pkl')

# PREPROCESSING
# No preprocessing is needed, because all the variables we are interested in are already in the appropriate format.

# ANALYSIS
## Table 1: "Logistic regression of diabetes status on physical activity, BMI, age, and their interaction terms, adjusting for sex, education, and income"
y = df_cleaned['Diabetes_binary']
X = df_cleaned[['PhysActivity', 'BMI', 'Age', 'Sex', 'Education', 'Income']]
X['PhysActivity_BMI'] = df_cleaned['PhysActivity'] * df_cleaned['BMI']
X['PhysActivity_Age'] = df_cleaned['PhysActivity'] * df_cleaned['Age']
```

```
X = sm.add_constant(X)
logit_model = sm.Logit(y, X)
result = logit_model.fit()

df1 = pd.DataFrame({
    'coef': result.params.values,
    'std err': result.bse,
    'pvalue': result.pvalues,
    '[0.025': result.conf_int()[0],
    '0.975]': result.conf_int()[1]},
    index=result.params.index)

df1.to_pickle('table_1.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
 'Total number of observations': len(df_cleaned),
 'Pseudo R-squ of regression model': result.prsquared
}
with open('additional_results.pkl', 'wb') as f:
    pickle.dump(additional_results, f)

```

C.2 Code Description

The code performs data analysis on a dataset of diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), specifically for the year 2015.

First, the dataset is loaded and any rows with missing or undefined values are removed. This cleaned dataset is then used for further analysis.

Descriptive statistics are calculated and saved in "table_0.pkl". These statistics include mean values of physical activity, BMI, age, and potential confounders (sex, education, and income), stratified by diabetes status.

Logistic regression analysis is performed to determine the relationship between diabetes status and physical activity, BMI, age, and their interaction terms. The model adjusts for potential confounders (sex, education, and income). The results of the regression analysis are saved in "table_1.pkl". The table includes coefficients, standard errors, p-values, and confidence intervals for each predictor variable.

Additionally, the code saves some additional results in "additional_results.pkl".

These results include the total number of observations in the cleaned dataset and the pseudo R-squared value of the regression model. The pseudo R-squared value provides an indication of the goodness of fit of the logistic regression model.

Overall, the code performs data analysis on the diabetes-related factors dataset, providing descriptive statistics and logistic regression results to gain insights into the relationship between diabetes status and various factors.

C.3 Code Output

table_0.pkl

	PhysActivity	BMI	Age	Sex	Education	Income
No Diabetes	0.777	27.79	7.814	0.4342	5.1	6.19
Diabetes	0.6306	31.93	9.379	0.4792	4.745	5.21

table_1.pkl

	coef	std err	pvalue	[0.025	0.975]
const	-4.348	0.07157	0	-4.488	-4.208
PhysActivity	-0.9705	0.08176	1.706e-32	-1.131	-0.8102
BMI	0.07675	0.001399	0	0.07401	0.07949
Age	0.1884	0.004106	0	0.1803	0.1964
Sex	0.3492	0.01248	2.417e-172	0.3247	0.3736
Education	-0.09969	0.006565	4.46e-52	-0.1126	-0.08683
Income	-0.1419	0.003175	0	-0.1481	-0.1357
PhysActivity_BMI	0.01375	0.001803	2.413e-14	0.01022	0.01728
PhysActivity_Age	0.02702	0.005013	7.083e-08	0.01719	0.03684

additional_results.pkl

```
{
    'Total number of observations': 253626,
    'Pseudo R-squ of regression model': 0.127             ,
}
```

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 typing import Dict, Tuple, Optional
from my_utils import to_latex_with_note, format_p_value

Mapping = Dict[str, Tuple[Optional[str], Optional[str]]]

# PREPARATION FOR ALL TABLES
def split_mapping(d: Mapping):
    abbrs_to_names = {abbr: name for abbr, (name, definition) in d.items() if name is not None}
    names_to_definitions = {name or abbr: definition for abbr, (name, definition) in d.items() if definition is not None}
    return abbrs_to_names, names_to_definitions

shared_mapping: Mapping = {
 'PhysActivity': ('Physical Activity', 'Physical Activity in past 30 days, 1: Yes, 0: No'),
 'BMI': ('BMI', 'Body Mass Index, kg/m2'),
 'Age': ('Age', '13-level age category in 5 years intervals'),
 'Sex': ('Sex', '0: Female, 1: Male'),
 'Education': ('Education Level', '1: Never attended school, 2: Elementary, 3: Some high school, 4: High school, 5: Some college, 6: College'),
 'Income': ('Income level', 'Income level on a scale of 1-8 (1: <=10K, 2: <=15K, 3: <=20K, 4: <=25K, 5: <=35K, 6: <=50K, 7: <=75K, 8: >75K)'),
}

# TABLE 0:
df = pd.read_pickle('table_0.pkl')
mapping = {k: v for k, v in shared_mapping.items() if k in df.columns or k in df.index}
abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(columns=abbrs_to_names, index=abbrs_to_names)

# Save as latex:
to_latex_with_note(
```

```
 df, 'table_0.tex',
 caption="Mean values of physical activity, BMI, age, and
 ↪ potential confounders stratified by diabetes status",
 label='table:descriptive_stats',
 legend=legend)


# TABLE 1:
df = pd.read_pickle('table_1.pkl')
df['pvalue'] = df['pvalue'].apply(format_p_value)

mapping = {k: v for k, v in shared_mapping.items() if k in
 ↪ df.index}
mapping |= {
 'pvalue': ('P-value', None),
 'coef': ('Coefficient', None),
 'PhysActivity_BMI': ('Physical Activity * BMI', 'Interaction
 ↪ term between Physical Activity and BMI'),
 'PhysActivity_Age':('Physical Activity * Age', 'Interaction
 ↪ term between Physical Activity and Age'),
 'const': ('Constant', None),
 'std err': ('Standard Error', None)
}

abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(index=abbrs_to_names)

# Save as latex:
to_latex_with_note(
 df, 'table_1.tex',
 caption="Logistic regression of diabetes status on physical
 ↪ activity, BMI, age, and their interaction terms,
 ↪ adjusting for sex, education, and income",
 label='table:logistic_regression',
 note="pvalue: If value is less than 1e-6, it is represented
 ↪ as less than 1e-6",
 legend=legend)

```

D.2 Code Output

table_0.tex

```
\begin{table}[h]
\caption{Mean values of physical activity, BMI, age, and potential confounders
    stratified by diabetes status}
\label{table:descriptive_stats}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrrrrrr}
\toprule
 & Physical Activity & BMI & Age & Sex & Education Level & Income level \\
\midrule
\textbf{No Diabetes} & 0.777 & 27.8 & 7.81 & 0.434 & 5.1 & 6.19 \\
\textbf{Diabetes} & 0.631 & 31.9 & 9.38 & 0.479 & 4.75 & 5.21 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Physical Activity}: Physical Activity in past 30 days, 1: Yes, 0:
    No
\item \textbf{BMI}: Body Mass Index, kg/m2
\item \textbf{Age}: 13-level age category in 5 years intervals
\item \textbf{Sex}: 0: Female, 1: Male
\item \textbf{Education Level}: 1: Never attended school, 2: Elementary, 3: Some
    high school, 4: High school, 5: Some college, 6: College
\item \textbf{Income level}: Income level on a scale of 1-8 (1: $<$=10K, 2:
    $<$=15K, 3: $<$=20K, 4: $<$=25K, 5: $<$=35K, 6: $<$=50K, 7: $<$=75K, 8: $>$75K)
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
\begin{table}[h]
\caption{Logistic regression of diabetes status on physical activity, BMI, age,
    and their interaction terms, adjusting for sex, education, and income}
\label{table:logistic_regression}
```

```
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrrlrr}
\toprule
 & coef & std err & pvalue & [0.025 & 0.975] \\
\midrule
\textbf{Constant} & -4.35 & 0.0716 & $<$1e-06 & -4.49 & -4.21 \\
\textbf{Physical Activity} & -0.97 & 0.0818 & $<$1e-06 & -1.13 & -0.81 \\
\textbf{BMI} & 0.0768 & 0.0014 & $<$1e-06 & 0.074 & 0.0795 \\
\textbf{Age} & 0.188 & 0.00411 & $<$1e-06 & 0.18 & 0.196 \\
\textbf{Sex} & 0.349 & 0.0125 & $<$1e-06 & 0.325 & 0.374 \\
\textbf{Education Level} & -0.0997 & 0.00657 & $<$1e-06 & -0.113 & -0.0868 \\
\textbf{Income level} & -0.142 & 0.00317 & $<$1e-06 & -0.148 & -0.136 \\
\textbf{Physical Activity * BMI} & 0.0137 & 0.0018 & $<$1e-06 & 0.0102 & 0.0173
    \\
\textbf{Physical Activity * Age} & 0.027 & 0.00501 & $<$1e-06 & 0.0172 & 0.0368
    \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item pvalue: If value is less than 1e-6, it is represented as less than 1e-6
\item \textbf{Physical Activity}: Physical Activity in past 30 days, 1: Yes, 0:
    No
\item \textbf{BMI}: Body Mass Index, kg/m2
\item \textbf{Age}: 13-level age category in 5 years intervals
\item \textbf{Sex}: 0: Female, 1: Male
\item \textbf{Education Level}: 1: Never attended school, 2: Elementary, 3: Some
    high school, 4: High school, 5: Some college, 6: College
\item \textbf{Income level}: Income level on a scale of 1-8 (1: $<$=10K, 2:
    $<$=15K, 3: $<$=20K, 4: $<$=25K, 5: $<$=35K, 6: $<$=50K, 7: $<$=75K, 8: $>$75K)
\item \textbf{Physical Activity * BMI}: Interaction term between Physical
    Activity and BMI
\item \textbf{Physical Activity * Age}: Interaction term between Physical
    Activity and Age
\end{tablenotes}
\end{threeparttable}
\end{table}
```

References

[1] 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.

[2] S. Akter, Md. Mizanur Rahman, Sarah Krull Abe, and P. Sultana. Prevalence of diabetes and prediabetes and their risk factors among bangladeshi adults: a nationwide survey. *Bulletin of the World Health Organization*, 92 3:204–13, 213A, 2014.

[3] K. Eckert. Impact of physical activity and bodyweight on health-related quality of life in people with type 2 diabetes. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 5:303 – 311, 2012.

[4] 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.

[5] Gitanjali M Singh, G. Danaei, F. Farzadfar, G. 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.

[6] Fanfan Zhao, Wentao Wu, Xiaojie Feng, Chengzhuo Li, Didi Han, Xiaojuan Guo, and Jun Lyu. Physical activity levels and diabetes prevalence in us adults: Findings from nhanes 20152016. *Diabetes Therapy*, 11:1303 – 1316, 2020.

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

[8] M. Yore, S. Ham, B. Ainsworth, J. Kruger, J. Reis, H. Kohl, and C. Macera. Reliability and validity of the instrument used in brfss to assess physical activity. *Medicine and science in sports and exercise*, 39 8:1267–74, 2007.

[9] W. Aekplakorn, S. Chariyalertsak, P. Kessomboon, S. Assanangkornchai, S. Taneepanichskul, and P. Putwatana. Prevalence of diabetes and relationship with socioeconomic status in the thai population: National health examination survey, 20042014. *Journal of Diabetes Research*, 2018, 2018.

[10] D. Umpierre, P. Ribeiro, C. Kramer, C. Leito, A. T. Zucatti, M. Azevedo, J. Gross, J. Ribeiro, and B. Schaan. Physical activity advice only or structured exercise training and association with hba1c levels in type 2 diabetes: a systematic review and meta-analysis. *JAMA*, 305 17:1790–9, 2011.

[11] Lei Zhang, L. Qin, Ai ping Liu, and Pei yu Wang. Prevalence of risk factors for cardiovascular disease and their associations with diet and physical activity in suburban beijing, china. *Journal of Epidemiology*, 20:237 – 243, 2010.