Association between BMI and the Prevalence of Diabetes

Data to Paper

July 2, 2023

Abstract

The association between Body Mass Index (BMI) and the prevalence of diabetes has significant implications for public health. Despite its global impact, the precise relationship between BMI and diabetes remains understudied. This research aims to bridge this gap by investigating the association between BMI and the prevalence of diabetes. We analyze a large-scale dataset derived from health indicators and employ a chi-squared test of independence. Our findings reveal a significant association between higher BMI and the prevalence of diabetes, supporting previous studies that highlight BMI as a risk factor for diabetes. Moreover, our results demonstrate that individuals with diabetes have higher mean BMI values compared to those without diabetes. These findings underscore the importance of BMI as a predictor for the prevalence of diabetes and emphasize the need for targeted interventions to address obesity. However, limitations of our study include reliance on self-reported data and potential confounding factors. Further research is necessary to understand the underlying mechanisms of this association, establish causality, and develop effective preventive and management strategies. By shedding light on the relationship between BMI and diabetes, this study contributes to the understanding of this complex health issue and provides valuable insights for public health policies and interventions.

Introduction

Diabetes, a disease of global concern [1], when exacerbated by the increasing prevalence of obesity, poses a significant public health challenge [2]. Among several risk factors associated with diabetes, Body Mass Index (BMI) is a crucial determinant representing a common thread among diabetes' disparate aspects [3, 4].

Existing research has provided crucial insights into the association between diabetes prevalence and BMI. Compelling evidence suggests marked disparities in diabetes risks among diverse demographic parameters, such as BMI [5]. Moreover, regional variations and gender-specific influences on diabetes risk among different BMI categories have been reported [6]. However, a more comprehensive understanding of this association can be facilitated by analyzing a broader range of BMI values at a population level.

In addressing this research gap, we leveraged health indicators in a rich dataset [7], earmarking data related to diabetes prevalence and demographic characteristics, inclusive of an extensive assortment of BMI values. The breadth and diversity inherent in this dataset afforded us the opportunity to conduct an intricate exploration into the association between BMI and diabetes at a granular level.

Our methodological approach involved a series of steps, including data preprocessing and the application of a chi-squared test of independence, a rigorously robust statistical test for investigating relationships between categorical variables [8]. Our findings indicated a significantly higher mean BMI among individuals with diabetes compared to those without. Furthermore, the chi-squared test underscored a notable association between diabetes and BMI, affirming the critical relevance of analyzing BMI for managing diabetes at a population level.

Results

The aim of our analysis was to investigate the association between Body Mass Index (BMI) and the prevalence of diabetes. To address this research question, we analyzed a large-scale dataset derived from health indicators. The dataset consisted of a total of 253,680 observations.

First, we examined the summary statistics for diabetes and BMI in the dataset (Table 1). Among individuals without diabetes, the mean BMI was 0.1838 (SD = 0.07316). In contrast, individuals with diabetes had a higher mean BMI of 0.2319 (SD = 0.08562). These summary statistics provide evidence of a higher BMI among individuals with diabetes compared to those without diabetes.

To test the association between diabetes and BMI, we conducted a chi-squared test of independence (Table 2). This statistical test was chosen because it allows us to examine the relationship between two categorical variables, in this case, diabetes (coded as ”Yes” and ”No”) and BMI. The chi-squared statistic was 12,942.93, indicating a significant association be-

Table 1: Summary statistics of diabetes and BMI in the dataset

	Diabetes	
	No	Yes
Count	218,334	35,346
Mean BMI	0.1838	0.2319
Std. Dev. of BMI	0.07316	0.08562
Min BMI	0	0.01163
25% percentile BMI	0.1395	0.1744
Median BMI	0.1744	0.2209
75% percentile BMI	0.2209	0.2674
Max BMI	1	1

tween diabetes and BMI ($p < 10^{-4}$). These findings support the presence of a strong relationship between BMI and the prevalence of diabetes in the dataset.

Table 2: Association between diabetes and BMI (Chi-squared test results)

Statistic	Value
Chi-squared statistic	12942.93
p-value	$< 10^{-4}$

In summary, our analysis revealed a significant association between BMI and the prevalence of diabetes. Individuals with diabetes had a higher mean BMI compared to individuals without diabetes. The chi-squared test results confirmed the presence of a significant association between these two variables. These findings highlight the importance of BMI as a risk factor for diabetes and suggest the need for targeted interventions to address obesity. However, it is important to acknowledge the limitations of our study, including the reliance on self-reported data and the potential for confounding factors. Further research is warranted to understand the underlying mechanisms of this association and establish causality.

Discussion

Our study was primarily focused on exploring the relationship between Body Mass Index (BMI) and the prevalence of diabetes, considering the underrepresentation of BMI in existing diabetes research literature [1]. The concern for diabetes on a global scale and its amplification alongside increasing obesity rates, championed the significance of this research [2].

Methodologically, we primarily used data preprocessing and a chi-squared test of independence to draw our conclusions. A chi-squared test of independence was favored as this test enables robust analysis of relationships between two categorical variables, and was deemed the most appropriate for our dataset [8].

The results of our study affirmed the conclusions of previous research, indicating a higher mean BMI among individuals with diabetes when compared to those without the disease [4, 6, 5]. Our chi-squared test reinforced the association between diabetes and BMI, aligning with past findings. However, the limitations of our study should be considered when interpreting these results. We relied on self-reported data, potentially leading to reporting bias. Confounding factors like physical activity and dietary patterns were also not controlled for, and could have influenced our observed associations [1]. Furthermore, our study design and cross-sectional data impede any verification of causality between BMI and the onset of diabetes.

Future research can conquer these limitations with methodologies that incorporate prospective cohort studies or randomized controlled trials, integrating a broader set of variables and facilitating a deeper understanding of this association.

In conclusion, our study emphasizes the significant association between BMI and the prevalence of diabetes, echoing the findings of previous research while accentuating the role of BMI as an important risk factor in diabetes. Our findings offer a foundation for future investigations and highlight the need for early preventive measures and interventions, particularly in the domain of obesity control [3]. Additionally, these results have substantial implications for public health, as the effective management of BMI could play a critical role in tackling the global health challenge posed by diabetes [1], warranting further exploration in future research.

Methods

Data Source

The data used in this study was obtained from the "diabetes_binary_health_indicators_BRFSS2015.csv" file. This dataset consists of various health indicators collected through the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey. The data includes information on diabetes status, high blood pressure, high cholesterol, cholesterol check, body mass index (BMI), smoking behavior, history of stroke and heart disease or heart attack, physical activity, fruit and vegetable consumption, heavy alcohol consumption, healthcare access, general health, mental health, physical health, difficulty with walking, sex, age, education, and income. Each row represents an individual record, and each column represents a specific health indicator or demographic characteristic.

Data Preprocessing

Prior to conducting the analysis, several preprocessing steps were performed on the dataset using Python. First, any missing values in the dataset were filled in using the mean value of the respective column. This approach ensured that all records contained complete information for further analysis.

Next, the BMI values were normalized using the MinMaxScaler from the scikit-learn library. This transformation rescaled the BMI values to a range between 0 and 1, allowing for a standardized comparison.

Data Analysis

To investigate the association between BMI and the prevalence of diabetes, a chi-squared test of independence was performed. First, the data was divided into groups based on BMI quartiles. This was achieved by applying the qcut function from the pandas library, which evenly partitioned the BMI values into four groups, ensuring an approximately equal number of observations in each group.

Next, a contingency table was created by cross-tabulating the diabetes status (binary) and the BMI groupings. This contingency table represented the observed frequencies of individuals with and without diabetes within each BMI group.

The chi-squared test of independence was then applied to the contingency table. This test determined whether there was a statistically significant association between BMI and diabetes. The test calculated the chi-squared statistic and the corresponding p-value. The chi-squared statistic measures the overall deviation from independence, while the p-value indicates the probability of obtaining such an extreme result, given the assumption of independence.

Finally, the results of the analysis were recorded in a results.txt file. This file included general information about the dataset, such as the total number of observations, the mean BMI value, and the percentage of subjects with diabetes. Additionally, two tables were created to summarize the results.

Table 1 presented descriptive statistics of BMI for individuals with and without diabetes, while Table 2 reported the chi-squared statistic and p-value resulting from the test of independence.

The code implementation provided a comprehensive analysis of the association between BMI and the prevalence of diabetes.

s

Code Availability

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

References

[1] A. Uloko, B. Musa, M. Ramalan, I. Gezawa, F. Puepet, A. Uloko, M. Borodo, and K. Sada. Prevalence and risk factors for diabetes mellitus in nigeria: A systematic review and meta-analysis. *Diabetes Therapy*, 9:1307 – 1316, 2018.

[2] A. Mokdad, B. Bowman, E. Ford, F. Vinicor, J. Marks, and J. Koplan. The continuing epidemics of obesity and diabetes in the united states. *JAMA*, 286 10:1195–200, 2001.

[3] Yeyi Zhu, M. Sidell, D. Arterburn, M. Daley, J. Desai, S. Fitzpatrick, M. Horberg, C. Koebnick, Emily V. McCormick, C. Oshiro, D. Young, and A. Ferrara. Racial/ethnic disparities in the prevalence of diabetes and prediabetes by bmi: Patient outcomes research to advance learning (portal) multisite cohort of adults in the u.s. *Diabetes Care*, 42:2211 – 2219, 2019.

[4] J. rnlv, J. Sundstrm, E. Ingelsson, and L. Lind. Impact of bmi and the metabolic syndrome on the risk of diabetes in middle-aged men. *Diabetes Care*, 34:61 – 65, 2010.

[5] S. Read, L. Rosella, H. Berger, D. Feig, K. Fleming, J. Ray, B. Shah, and L. Lipscombe. Bmi and risk of gestational diabetes among women of south asian and chinese ethnicity: a population-based study. *Diabetologia*, 64:805 – 813, 2021.

[6] Y. Rho, N. Lu, C. Peloquin, Ada Man, Yanyan Zhu, Yuqing Zhang, and Hyon K. Choi. Independent impact of gout on the risk of diabetes mel-

litus among women and men: a population-based, bmi-matched cohort study. *Annals of the Rheumatic Diseases*, 75:91 – 95, 2014.

[7] Henock M. Deberneh and Intaek Kim. Prediction of type 2 diabetes based on machine learning algorithm. *International Journal of Environmental Research and Public Health*, 18, 2021.

[8] Morgana Mongraw-Chaffin, S. Peters, R. Huxley, and M. Woodward. The sex-specific association between bmi and coronary heart disease: a systematic review and meta-analysis of 95 cohorts with 12 million participants. *The lancet. Diabetes & endocrinology*, 3 6:437–449, 2015.

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)
```

Data Exploration

Code

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

```python

import pandas as pd
import numpy as np

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

# Open the output file for writing
output_file = open("data_exploration.txt", "w")

# Measure of the scale of the data
num_rows = len(df)
num_cols = len(df.columns)
```

```python
output_file.write("Number of rows: {}\n".format(num_rows))
output_file.write("Number of columns:
↪  {}\n\n".format(num_cols))

# Summary statistics of key variables
output_file.write("Summary statistics:\n")
output_file.write(df.describe().to_string() + "\n\n")

# List of most common values of categorical variables
output_file.write("Most common values of categorical
↪  variables:\n")
for column in df.select_dtypes(include=['object']):
    common_values = df[column].value_counts().reset_index()
    common_values.columns = [column, 'Count']
    output_file.write("{}:\n{}\n\n".format(column,
    ↪  common_values))

# Counts of missing, unknown, or undefined values
output_file.write("Counts of missing, unknown, or undefined
↪  values:\n")
for column in df.columns:
    unknown_count = df[column].isin(['unknown',
    ↪  'undefined']).sum()
    missing_count = df[column].isna().sum()
    output_file.write("{} - Unknown values: {}, Missing
    ↪  values: {}, Total: {}\n".format(column, unknown_count,
    ↪  missing_count, num_rows))

output_file.close()

```

Code Description

The Data Exploration code is designed to analyze and summarize the given dataset, specifically focusing on key variables and identifying missing, unknown, or undefined values. This code is useful for gaining insights into the dataset's characteristics and preparing it for further analysis or modeling.

The code begins by loading the dataset from the file ”diabetes_binary_health_indicators_BRFSS2015. using pandas library. It then opens a new file, ”data_exploration.txt”, for writing the analysis results.

The code first calculates and writes the number of rows and columns in the dataset, providing a measure of the dataset’s scale.

Next, the code computes and writes the summary statistics for the numerical variables in the dataset. This includes statistical measures such as count, mean, standard deviation, minimum, and maximum values.

Following that, the code identifies and writes the most common values for each categorical variable. It counts the occurrence of each unique value and presents them in descending order of frequency.

Lastly, the code determines and writes the counts of missing, unknown, or undefined values for each variable in the dataset. It identifies missing values using the isna() function, and unknown or undefined values based on specific keywords such as ’unknown’ or ’undefined’. These counts provide information about the data quality and potential issues with missing or unreliable data.

The output of the code is written into the ”data_exploration.txt” file. The file contains the following information: - Number of rows and columns in the dataset - Summary statistics for the numerical variables - Most common values for each categorical variable - Counts of missing, unknown, or undefined values for each variable

This file serves as a record of the data exploration process and can be used to understand the dataset better and make informed decisions in subsequent data analysis steps.

Code Output

```
Number of rows: 253680
Number of columns: 22

Summary statistics:
                      count    mean    std
Diabetes_binary      253680  0.1393 0.3463
HighBP               253680   0.429 0.4949
HighChol             253680  0.4241 0.4942
CholCheck            253680  0.9627 0.1896
BMI                  253680   28.38  6.609
Smoker               253680  0.4432 0.4968
Stroke               253680 0.04057 0.1973
HeartDiseaseorAttack 253680 0.09419 0.2921
PhysActivity         253680  0.7565 0.4292
Fruits               253680  0.6343 0.4816
```

```
Veggies              253680  0.8114 0.3912
HvyAlcoholConsump    253680  0.0562 0.2303
AnyHealthcare        253680  0.9511 0.2158
NoDocbcCost          253680 0.08418 0.2777
GenHlth              253680   2.511  1.068
MentHlth             253680   3.185  7.413
PhysHlth             253680   4.242  8.718
DiffWalk             253680  0.1682 0.3741
Sex                  253680  0.4403 0.4964
Age                  253680   8.032  3.054
Education            253680    5.05 0.9858
Income               253680   6.054  2.071

Most common values of categorical variables:
Counts of missing, unknown, or undefined values:
Diabetes_binary - Unknown values: 0, Missing values: 0, Total: 253680
HighBP - Unknown values: 0, Missing values: 0, Total: 253680
HighChol - Unknown values: 0, Missing values: 0, Total: 253680
CholCheck - Unknown values: 0, Missing values: 0, Total: 253680
BMI - Unknown values: 0, Missing values: 0, Total: 253680
Smoker - Unknown values: 0, Missing values: 0, Total: 253680
Stroke - Unknown values: 0, Missing values: 0, Total: 253680
HeartDiseaseorAttack - Unknown values: 0, Missing values: 0, Total: 253680
PhysActivity - Unknown values: 0, Missing values: 0, Total: 253680
Fruits - Unknown values: 0, Missing values: 0, Total: 253680
Veggies - Unknown values: 0, Missing values: 0, Total: 253680
HvyAlcoholConsump - Unknown values: 0, Missing values: 0, Total: 253680
AnyHealthcare - Unknown values: 0, Missing values: 0, Total: 253680
NoDocbcCost - Unknown values: 0, Missing values: 0, Total: 253680
GenHlth - Unknown values: 0, Missing values: 0, Total: 253680
MentHlth - Unknown values: 0, Missing values: 0, Total: 253680
PhysHlth - Unknown values: 0, Missing values: 0, Total: 253680
DiffWalk - Unknown values: 0, Missing values: 0, Total: 253680
Sex - Unknown values: 0, Missing values: 0, Total: 253680
Age - Unknown values: 0, Missing values: 0, Total: 253680
Education - Unknown values: 0, Missing values: 0, Total: 253680
Income - Unknown values: 0, Missing values: 0, Total: 253680
```

Data Analysis

Code

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

```python

import pandas as pd
from scipy.stats import chi2_contingency
from sklearn.preprocessing import MinMaxScaler

# (1) Load the data
data = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.csv")

# (3) Preprocessing
# Deal with missing, unknown, or undefined values
data = data.fillna(data.mean())

# Normalization of BMI
MMScaler = MinMaxScaler()
data['BMI'] = MMScaler.fit_transform(data[['BMI']])

# (4) Perform the analysis
# We'll make groups based on BMI quartiles
data['BMI_group'] = pd.qcut(data['BMI'], q=4)

contingency = pd.crosstab(data['Diabetes_binary'], data['BMI_group'])

# Chi-squared test of independence
chi2, p_value, _, _ = chi2_contingency(contingency)

# (5) Create and output the data analysis results
with open("results.txt", "w") as file:
    # General results
    file.write(f"## General results:\n")
    file.write(f"\nTotal number of observations: {data.shape[0]}\n")
    file.write(f"\nMean BMI: {data['BMI'].mean()}\n")
    file.write(f"\nPercentage of subjects with diabetes:
    ↪  {data['Diabetes_binary'].mean()*100}\n")
    file.write("\n")

    # Results for Table 1
    file.write(f"## Results for Table 1:\n")
    diabetes_bmi_summary =
    ↪  data.groupby('Diabetes_binary')['BMI'].describe().transpose()
    file.write(f"\n{diabetes_bmi_summary}\n")

    # Results for Table 2
    file.write(f"## Results for Table 2:\n")
    file.write(f"\nChi-squared statistic: {chi2}\n")
    file.write(f"\np-value: {p_value}\n")

```

Code Description

The given code performs a data analysis on a dataset containing health indicators related to diabetes. The main objective is to explore the relationship between diabetes and body mass index (BMI), and to generate some analysis results.

Firstly, the code loads the dataset into a pandas DataFrame. It then proceeds with preprocessing steps. Missing values in the dataset are filled with the mean of their respective columns. Additionally, the BMI values are normalized using a MinMaxScaler.

Next, the code performs the analysis by making groups based on quartiles of BMI. It creates a contingency table by cross-tabulating the diabetes status and the BMI groups. This contingency table is then used to perform a chi-squared test of independence, which helps determine if there is a statistically significant relationship between diabetes and BMI.

The analysis results are written into a file called ”results.txt”. The file begins with some general results, including the total number of observations, the mean BMI value, and the percentage of subjects with diabetes.

Following the general results, the code generates two tables. Table 1 provides a summary of BMI statistics (mean, standard deviation, minimum, maximum, quartiles) for both diabetes and non-diabetes groups. Table 2 presents the chi-squared statistic and the p-value resulting from the chi-squared test, indicating the strength and significance of the relationship between diabetes and BMI.

These output files can be used to inform further analysis or to present the findings of the study.

Code Output

```
## General results:

Total number of observations: 253680

Mean BMI: 0.19049260008947363

Percentage of subjects with diabetes: 13.933301797540206

## Results for Table 1:

Diabetes_binary       0       1
count            218334   35346
mean             0.1838  0.2319
std             0.07316 0.08562
min                   0 0.01163
25%              0.1395  0.1744
50%              0.1744  0.2209
75%              0.2209  0.2674
max                   1       1
## Results for Table 2:

Chi-squared statistic: 12942.93491586499

p-value: 0.0
```