Introduction

According to the American Diabetes Association, approximately 1.2 million Americans are diagnosed annually. By 2021, about 11.6% of the American population had diabetes (Statistics About Diabetes | ADA). The International Diabetes Federation (IDF) predicts that by 2045, 1 in 8 adults will be living with diabetes, an increase of 46%, highlighting its status as a major public health concern not only in the United States but worldwide (International Diabetes Federation). Given the high prevalence of diabetes and the IDF's alarming predictions, researchers have worked to identify causes, prevention, and management of the disease.

While genetics also plays a role, behavioral factors have emerged as the focus of research to understanding diabetes risk. The American Psychological Association notes that the development and course of diabetes is intensely related to multiple behavioral factors, mentioning dietary habits, physical activity, medication adherence, and blood glucose monitoring (Hunter et al. 2024). Examining how these factors interact with diabetes diagnosis may lead to further research opportunities and how the disorder is viewed and managed.

Research supports the idea that behavioral and lifestyle factors not only influence the development of diabetes but also play an essential role in its management. For example, current research suggests that lifestyle changes, such as physical activity and dietary adjustments, can mitigate diabetes risk and improve patient outcomes. However, there remains a need for further study to clarify connections between patient outcomes and sustainable healthy habits, particularly in identifying additional predictors, such as educational level, family history, or financial status, that can expand preventative measures (Calhoun et al. 2019). Although lifestyle changes have been identified as key to preventing and managing diabetes, there is a gap in understanding which specific lifestyle and medical factors most effectively predict a diabetes diagnosis. Addressing this need could lead to actionable changes, such as improvement in healthcare policies or personal interventions.

The goal of this research is to answer the question: What lifestyle and medical factors are most effective in predicting a diabetes diagnosis? While medical factors have led to current diabetes risk prevention and diabetes management, lifestyle factors require further investigation to possibly expand the approach. By situating this study within the greater context of diabetes research and focusing on modifiable lifestyle and medical factors empirically chosen, such as BMI, physical activity, and age, this work aims to inform the development of prevention strategies and enhance diabetes management.

Methods

Our study uses the Diabetes Health Indicators Dataset, collected by the Centers for Disease Control and Prevention (CDC) through the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS data set is one of the largest global health surveys that are ongoing, designed to collect data on health-related risk behaviors, chronic health conditions, and the use of preventative services in adults aged 18 and older. The survey uses random digit dialing to capture a wide range of individuals, using both landlines and cell phones. This specific survey gathered responses from over 400,000 Americans in 2015. This survey includes 441,455 individuals and 330 variables. The Diabetes Health Indicators Dataset narrows this focus to diabetes-relevant variables. This subset, pre-cleaned to exclude missing values, consists of 253,680 individuals and 22 variables.

The goal of this research is to consider various variables as predictors of diabetes diagnosis. Our dataset includes diabetes diagnosis as a binary variable, coded as:

- 0 for individuals without diabetes
- 1 for individuals with either diabetes or prediabetes

This binary coding serves as the response variable in the analysis.

The final model incorporates seven explanatory variables, consisting of one continuous variable, two categorical variables, and four binary variables. These were chosen based on the theoretical relevance and documented research with diabetes in previous studies, which will be explained more in-depth later. The variables are:

1. **BMI (Body Mass Index):** A numeric variable being measured using kilograms per meter squared (kg/m^2). The original data set recorded BMI values ranging from 12 to 98; however, when outliers are removed from the data set, the range narrowed to 14 to 41.
2. **High Blood Pressure:** A binary variable coded as 1 if the individual has been told by a healthcare professional that they have high blood pressure and 0 otherwise.
3. **High Cholesterol:** A binary variable coded as a 1 if an individual had their cholesterol checked and had been told by a doctor, nurse, or other health professional that it was high and 0 otherwise.
4. **Physical Activity:** A binary variable coded as 1 if the individual reported having had physical activity in the last 30 days excluding one's job, and 0 otherwise.

5. **Heavy Alcohol Consumption:** A binary variable coded as a 1 if the individual indulged in heavy alcohol consumption (more than or equal to 14 drinks per week for men and more than or equal to 7 drinks per week for women) and 0 otherwise.
6. **Age:** A categorical variable being represented with numbers ranging from 1 to 13. The age groups start from ages 18 to 24 (coded as 1) and end with ages 80 and older (coded as 13). A detailed breakdown can be found in *Figure 1*.
7. **Income:** A categorical variable represented by integers ranging from 1 to 8, starting with less than $10,000 a year (coded as 1) to $75,000 or more a year (coded as 8). A detailed breakdown can be found in *Figure 1*.

Variable	Coding
Age	1 = Age 18 to 24
	2 = Age 25 to 29
	3 = Age 30 to 34
	4 = Age 35 to 39
	5 = Age 40 to 44
	6 = Age 45 to 49
	7 = Age 50 to 54
	8 = Age 55 to 59
	9 = Age 60 to 64
	10 = Age 65 to 69
	11 = Age 70 to 74
	12 = Age 75 to 79
	13 = Age 80 or older
Income	1 = Less than $10,000
	2 = $10,000 to less than $15,000
	3 = $15,000 to less than $20,000
	4 = $20,000 to less than $25,000
	5 = $25,000 to less than $35,000
	6 = $35,000 to less than $50,000
	7 = $50,000 to less than $75,000
	8 = $75,000 or more

Figure 1: Coding for Income and Age

This selection includes a mix of numeric, binary, and categorical variables, which cover a wide range of possible factors in diabetes diagnosis. We also explored two additional binary variables that were further removed from our modeling: fruits (if an individual eats fruit one or more times a day) and vegetables (if an individual eats vegetables one or more times a day). Our decision on why we excluded these variables will be explored in our results section.

The data set required minimal cleaning. As mentioned previously, BMI included a large range of values despite the upper half of the range not only being outliers, but also not making sense to the application since morbid obesity begins at 40. We decided to remove all outliers (defined as any value below the first quartile by 1.5 times the inner quartile range or above the third quartile by 1.5 times the inner quartile range) for BMI from the data set to avoid potential data entry errors. This decreased our data set to 243,833 individuals, removing 9847 entries. All other variables did not require such actions, nor did any variable have any missing values to address. The cleaned dataset ensured robust and reliable modeling of diabetes diagnosis predictors.

Modeling

To prepare for modeling, we investigated our variables. The initial data set included 22 variables spanning medical, lifestyle, and general health factors. Overly generalized variables (e.g. general health and mental health) were excluded to focus on specific lifestyle predictors. While two medical factors directly related to diabetes risk (high blood pressure and high cholesterol) were retained, we decided to exclude the majority to align with our focus on lifestyle factors rather than medical factors. This refinement process reduced our focus to a subset of nine variables for exploratory data analysis (EDA).

Modeling began with an initial model including all nine variables and a reduced model including variables that were highlighted in the exploratory data analysis (BMI, physical activity, heavy alcohol consumption, and age). These models were evaluated using likelihood ratio tests, Bayesian Information Criterion (BIC), and misclassification rates. The full model demonstrated lower BIC values indicating improved performance despite the complexity, while the misclassification rates were extremely similar (<1% difference).

Model	BIC	Misclassification Rate (%)
Full (8 vars)	169484.8	13.78%
Full + Interaction (BMI x PhysAct)	169454.1	13.77%
EDA (4 vars)	181373.8	14.10%
EDA + Interaction (BMI x PhysAct)	181276.8	14.09%
Final (6 vars)	169536.2	13.78%

Figure 2: Tested models and their Bayesian Information Criterion (BIC) and misclassification rate (%).

Interaction terms between BMI and physical activity were tested in both models due to their suspected correlation. While the interaction models slightly improved performance metrics, the improvement in BIC and misclassification rates were negligible, leading to their exclusion from the final model. Other interactions were explored, however, yielded weaker similar results.

Stepwise regression confirmed the removal of fruits and vegetables, as their exclusion had minimal impact on BIC. Further removal was halted after physical activity was identified as the next candidate, as it remained significant and relevant to our research according to our exploratory data analysis. The final model included six variables: BMI, age, income, high blood pressure, high cholesterol, physical activity, and heavy alcohol consumption.

Our final model was assessed for adherence to logistic regression conditions. Linearity was confirmed using an empirical logit plot for our single numeric variable, BMI, and confirmed a linear relationship with the log odds of diabetes diagnosis. This justifies our decision to ignore any data transformations. The BRFSS sampling design confirms the randomness and independence conditions for all variables. While multicollinearity was explored, there were no correlations between exploratory variables that seemed significant enough to prevent interpretation.

The final data set included 243,833 individuals. BMI showed a mean of 27.57 kg/m² and a range of 14 to 41 kg/m² after outlier removal, with most values clustering between 24 and 31 kg/m². This distribution reflects a mix of underweight, normal-weight, overweight, and obese individuals in the dataset. About 75.7% of individuals in the dataset reported having had physical activity outside of work within the last 30 days, and 42.9% had been diagnosed with high blood pressure. High cholesterol was present in 42.4% of the individuals and heavy alcohol consumption was present in 5.6% of the individuals. Age was generally normally distributed with the largest group of 13.1% individuals ages 60 to 64 and our smallest group of 2.25% individuals ages 18 to 24. Income was skewed left with the largest group of 35.6% individuals with an income of $75,000 or larger and the smallest group of 3.87% individuals with an income of less than $10,000.

Terms	**CI for OR (95%)**	**OR**	**β**	**Standard Error**	**Wald Test Results (p-value)**
Intercept	(0.0010, 0.0017)	0.0013	-6.637	0.1232	< 2e-16 ***
Income ≥ $10,000 & < $15,000	(0.8895, 1.0280)	0.9562	-0.0448	0.0348	0.2255
Income ≥ $15,000 & < $20,000	(0.7714, 0.8863)	0.8268	-0.1902	0.0333	7.88e-08 ***
Income ≥ $20,000 & < $25,000	(0.7083, 0.8103)	0.7575	-0.2777	0.0324	6.03e-16 ***
Income ≥ $25,000 & < $35,000	(0.6056, 0.6906)	0.6466	-0.4360	0.0316	< 2e-16 ***
Income ≥ $35,000 & < $50,000	(0.5168, 0.5870)	0.5507	-0.5965	0.0307	< 2e-16 ***
Income ≥ $50,000 & < $75,000	(0.4748, 0.5389)	0.5058	-0.6816	0.0305	< 2e-16 ***
Income ≥ $75,000	(0.3790, 0.4279)	0.4026	-0.9097	0.0293	< 2e-16 ***

Figure 3.1: The 95% confidence intervals for odd ratios, coefficients calculated for the model, odds ratio calculated from the coefficients, and results from the Wald Test for our logistic regression model.

**** Coefficients were extremely significantly different from 0 at* $\alpha = 0.005$.

OR: Odds Ratio CI: Confidence Interval β: Coefficient of term in model

Terms	CI for OR (95%)	OR	β	Standard Error	Wald Test Results (p-value)
High Blood Pressure	(2.3644, 2.5069)	2.4345	0.8898	0.0144	< 2e-16 ***
High Cholesterol	(1.8328, 1.9348)	1.883	0.6329	0.0132	< 2e-16 ***
Physical Activity	(0.7689, 0.8130)	0.7907	-0.2349	0.0136	< 2e-16 ***
Heavy Alcohol Consumption	(0.4206, 0.4900)	0.4543	-0.7891	0.0378	< 2e-16 ***
BMI	(1.1049, 1.111)	1.1078	0.1024	0.0009	< 2e-16 ***
Ages 25 to 29	(0.7537, 1.3886)	1.0200	0.0198	0.1460	0.8985
Ages 30 to 34	(1.1452, 1.9619)	1.4896	0.3985	0.1310	0.0036 **
Ages 35 to 39	(1.7645, 2.9294)	2.2565	0.8138	0.1243	2.77e-10 ***
Ages 40 to 44	(2.3488, 3.8538)	2.9848	1.0935	0.1216	< 2e-16 ***
Ages 45 to 49	(3.0580, 4.9768)	3.8691	1.3530	0.1199	< 2e-16 ***
Ages 50 to 54	(3.8573, 6.2441)	4.8663	1.5823	0.1187	< 2e-16 ***
Ages 55 to 59	(4.2615, 6.8861)	5.3712	1.6810	0.1183	< 2e-16 ***
Ages 60 to 64	(5.2949, 8.5455)	6.6692	1.8975	0.1180	< 2e-16 ***
Ages 65 to 69	(6.1122, 9.8616)	7.6974	2.0409	0.1180	< 2e-16 ***
Ages 70 to 74	(6.5208, 10.5322)	8.2167	2.1068	0.1183	< 2e-16 ***
Ages 75 to 79	(6.2687, 10.1485)	7.9089	2.0680	0.1189	< 2e-16 ***
Ages 80 or older	(5.8102, 9.4069)	7.3307	1.9921	0.1190	< 2e-16 ***

Figure 3.2: The 95% confidence intervals for odd ratios, coefficients calculated for the model, odds ratio calculated from the coefficients, and results from the Wald Test for our logistic regression model.

*** Coefficients were significantly different from 0 at* $\alpha = 0.05$.

**** Coefficients were extremely significantly different from 0 at* $\alpha = 0.005$.

OR: Odds Ratio CI: Confidence Interval β: Coefficient of term in model

Figures 3.1 and 3.2 present the parameter estimates, standard errors, p-values from the Wald tests, odds ratios (OR), and 95% confidence intervals (CIs) for the odds ratios for each variable in the

final logistic regression model. In Figure 3.1, the odds of an individual with an income larger than $75,000 having diabetes were 0.402x the odds of an individual with an income less than $10,000 having diabetes (95% CI: [0.38, 0.43], p < 2e-16), suggesting that having a lower income can predict diabetes risk. The 95% CI does not include 1, further confirming the importance of this relationship. In Figure 3.2, high blood pressure's odds ratio indicates that the odds of having high blood pressure and diabetes were 2.43x the odds of not having high blood pressure and having diabetes (95% CI: [2.36, 2.51], p < 2e-16). The 95% CI excluding 1 further suggests that this is a strong predictor of diabetes risk. Heavy alcohol consumption was identified as another strong indicator of diabetes diagnosis, with the odds of an individual with heavy alcohol consumption having diabetes being 0.45x the odds of an individual without heavy alcohol consumption having diabetes (95% CI: [0.42, 0.49], p < 2e-16). The 95% CI does not include 1, further solidifying that alcohol consumption is a strong predictor of diabetes diagnosis. For age, the odds of an individual aged 70 to 74 having diabetes were 8.2x the odds of an individual aged 18 to 24 having diabetes (95% CI: [6.52, 10.53], p < 2e-16). The 95% CI does not include 1, suggesting that age is a very strong predictor of diabetes risk. The odds of an individual with high cholesterol having diabetes were 1.88x the odds of an individual without high cholesterol having diabetes (95% CI: [1.83, 1.93], p < 2e-16). The 95% CI excluding 1 indicates that this relationship is significant. The odds of an individual who had physical activity having diabetes were 0.79x the odds of an individual who did not have physical activity having diabetes (95% CI: [0.77, 0.81], p < 2e-16). The 95% CI excludes 1, which indicates that this is an important relationship. BMI showed a relatively weak relationship to diabetes diagnosis, with the 95% confidence internal almost including 1, with the lower bound of 1.105 (95% CI: [1.105, 1.11], p < 2e-16). For a one unit increase in BMI, the odds of having diabetes were multiplied by 1.11x the odds of not having diabetes.

The strongest predictors of diabetes diagnosis in our model were income, high blood pressure, heavy alcohol consumption, and age, however, all the terms in the model were predictive. High blood pressure, high cholesterol, and age were associated with increased odds of diabetes, while physical activity, heavy alcohol consumption, and income were associated with a decrease in odds of diabetes. Overall, these results support the hypothesis of past research that physical activeness and retaining healthy blood pressure and cholesterol is important for diabetes risk prevention, with additional factors being highlighted.

The final model showed good fit with significant predictors that align with established diabetes risk factors. The model's performance metrics support its validity for understanding diabetes predictors.

Discussion

As hypothesized according to past research, age, high blood pressure, and high cholesterol are associated with an increased likelihood of diabetes, while physical activity is associated with a decreased likelihood of diabetes. Additionally, we found that heavy alcohol consumption and income can also predict diabetes diagnosis with a decreased likelihood of diabetes.

The result for heavy alcohol consumption is unexpected, as previous research generally suggests a positive relationship between alcohol consumption and diabetes risk. There are a few explanations for why this may have occurred. One possible explanation for this finding could be from additional unmeasured factors that may be influencing the relationship, such as diet, medication, or access to healthcare. Alternatively, there may have been the presence of reverse causality, where individuals diagnosed with diabetes may reduce or stop heavy drinking as part of lifestyle changes after diagnosis.

The strong association of income and diabetes highlights the need for social work programs to support low-income individuals accessing diabetes risk prevention information and programs. Additionally, age's strong association with diabetes indicates that advocating for risk prevention must increase as individuals age, as they are more at risk than when they were younger. The associations of high blood pressure, high cholesterol, and physical activity serve as a reminder that taking care of one's body not only helps those regions of health, but also prevents and manages diabetes.

The BRFSS dataset allows us to generalize these results to the broader United States adult population; however, the findings may not fully apply to populations outside of the U.S. or to subgroups underrepresented in the data set, such as younger individuals. A limitation of this study is the reliance on this self-reported data, specifically with alcohol consumption, which could introduce recall bias or social desirability bias. There may be additional confounding variables that may have affected our results, such as genetic predisposition, education, and ethnicity. These factors could influence the observed relationships between our lifestyle variables and diabetes diagnosis and warrant further investigation. A strength of this analysis is the use of a large, nationally representative data set, as well as the data subset that was created before the start of our research that was specifically designated to be for diabetes analysis. However, the lack of interaction terms and the reliance on a national dataset may limit the scope of our conclusions.

Future studies should aim to dissect the relationship between alcohol consumption and diabetes more thoroughly, possibly investigating using longitudinal data to better understand causality, as well as exploring how the relationship shifts by population subgroups. Investigating the impact of healthcare access and socioeconomic factors on diabetes risk and prevalence could offer valuable insights.

References

Calhoun, Nicole, et al. “Are Demographic Factors Associated with Diabetes Risk Perception and Preventive Behavior?” *Journal of Best Practices in Health Professions Diversity*, vol. 12, no. 2, 2019, pp. 128–40. *JSTOR*, https://www.jstor.org/stable/26954204. Accessed 12 Nov. 2024.

“Diabetes in Older People.” *National Institute on Aging*, 10 Apr. 2024, www.nia.nih.gov/health/diabetes/diabetes-older-people.

Gray, Natallia, et al. “Relation Between BMI and Diabetes Mellitus and Its Complications Among US Older Adults.” *Southern Medical Journal*, vol. 108, no. 1, Jan. 2015, pp. 29 36. https://doi.org/10.14423/smj.0000000000000214.

Hunter, Christopher L, et al. *Integrated Behavioral Health in Primary Care*. American Psychological Association (APA), 2 Jan. 2024.

International Diabetes Federation. “Diabetes Facts and Figures | International Diabetes Federation.” *International Diabetes Federation*, 7 May 2024, idf.org/about-diabetes/diabetes-facts-figures.

Kim, Soo-Jeong, and Dai-Jin Kim. “Alcoholism and Diabetes Mellitus.” *Diabetes & Metabolism Journal*, vol. 36, no. 2, Jan. 2012, p. 108. https://doi.org/10.4093/dmj.2012.36.2.108.

Statistics About Diabetes | ADA. diabetes.org/about-diabetes/statistics/about-diabetes.

Zahalka, Salwa J., et al. “The Role of Exercise in Diabetes.” *Endotext - NCBI Bookshelf*, 6 Jan. 2023, www.ncbi.nlm.nih.gov/books/NBK549946.