Modeling the Determinants of Diabetes and Prediabetes Using the 2015 CDC Behavioral Risk Factor Surveillance System (BRFSS) Data

Oluwafunmibi Omotayo Fasanya
Louisiana State University Health Sciences Center, New Orleans
Categorical Data Analysis Project (BSDS 6210)

Background and analysis objectives

Diabetes is a serious chronic health condition that occurs when the blood sugar (glucose) is high because the body has lost the ability to process and use food effectively. When food is consumed, it is broken down into glucose during digestion and then released into the bloodstream which signals the pancreas to produce the hormone insulin that allows the conversion of the food into usable energy. However, in individuals with diabetes, the process is hindered as the pancreases does not produce enough insulin, or the body is unable to use it properly. As a result, glucose remains in the bloodstream instead of being absorbed by the cells for energy. This health condition poses a serious threat to health in the 21st century. According to (Winer & Sowers, 2013), with diabetes comes an increased risk of stroke, heart/kidney failure, birth complications, sexual dysfunction, limb amputations, and the common risk factors associated with diabetes include obesity, hypertension and dyslipidemia.

This disease is a growing global health issue that affects about 463 million adults worldwide and is projected to rise to 700 million by 2045 (Gao, 2025). According to the center for Disease Control and Prevention (CDC), diabetes is part of the leading top 10 causes of death in the United States, about 38.4 million Americans (11.6% of the United State population) currently have diabetes, 29.7 million have been diagnosed, while 8.7 million (22.8%) are unaware that they have diabetes. Also, 97.6 million people aged 18 years or older (38% of the United State population) have prediabetes, a condition that significantly increases the risk of developing type 2 diabetes. Of the total prevalence of Prediabetes, approximately 27.2 million (48.8%) of older adults aged 65 years and above are living with this condition (Centers for Disease Control and Prevention [CDC], 2024).

This project aims to analyze how demographic characteristics, health behaviors, mental and physical health status, cardiovascular comorbidities and healthcare access contributes to the risk of diabetes and prediabetes using the Centers for Disease Control and Prevention 2015 Behavioral Risk Factor Surveillance System survey dataset.

Data description and data cleaning

Data Description

The diabetes health indicators dataset that was used for this project was obtained from Kaggle. This dataset is originally gotten from the CDC'S 2015 Behavioral Risk Factor Surveillance System (BRFSS) but has been cleaned. The BRFSS is a cross-sectional telephone survey conducted by the Centers for Disease Control and Prevention and state health departments on over 400, 000 Americans to collect data on health-related risk behaviors, chronic health conditions, and the use of preventative services. The 2015 BRFSS dataset has 441,455 individuals and 330 features, these features contain questions in the survey and others that were automatically generated based on respondents' response. This project used the diabetes _ 012 _ health _ indicators _ BRFSS2015 dataset. It is a clean dataset of 253,680 individual responses and includes 21 features variables with the target variable Diabetes_012 which has 3 classes. 0 is for no diabetes or only during pregnancy, 1 for prediabetes, and 2 for diabetes. There is class imbalance in this dataset, and the predictor variables are listed below (Table 1)

Data Cleaning

The dataset was already cleaned and has no missing variable. However, factor encoding was done on the categorical independent and dependent variables, appropriate names were given to the levels and variables were converted to factor and ordinal as appropriate. The response variable has three levels namely no diabetes or only during pregnancy, prediabetes, and diabetes, we created another variable called diabetes_binary that has no diabetes or only during pregnancy as a level and both prediabetes, and diabetes as another level.

Exploratory data analysis through numeric summary and/or graphs

Table 2 shows the demographic characteristics (Sex, age, education and household income) of respondents stratified by diabetes status (No diabetes, Prediabetes, and Diabetes). Each of these variables are summarized using counts and percentage. A Pearson's Chi-square test was conducted to

TABLE 1. List of Variables Included in the Analysis

Category	**Variables and Descriptions**
Demographic	Sex: Male, Female Age: 14 categories (18–24, 25–29, ..., 80+) Education: 6 levels (Never attended school, Elementary, Some high school, High school graduate, Some college, College graduate) Income: 8 categories ($<10k to $75k+)
Body Mass Index	BMI: Body Mass Index
High Blood Pressure	HighBP: Told by a health professional you have high blood pressure (Yes/No)
High Cholesterol	HighChol: Told cholesterol is high (Yes/No) CholCheck: Cholesterol check in past 5 years (Checked / Not checked)
Smoking	Smoker: Ever smoked at least 100 cigarettes (Yes/No)
General and Mental Health	GenHlth: General health (Excellent, Very good, Good, Fair, Poor) MentHlth: Days mental health not good (past 30 days) PhysHlth: Days physical health not good (past 30 days) DiffWalk: Difficulty walking or climbing stairs (Yes/No)
Chronic Health Conditions	Stroke: Ever told you had a stroke (Yes/No) HeartDiseaseorAttack: Ever told you had CHD or MI (Yes/No)
Health Care Access	AnyHealthcare: Has healthcare coverage (Yes/No) NoDocbcCost: Could not see a doctor due to cost (Yes/No)
Physical Activity	PhysActivity: Any physical activity in last 30 days (Yes/No)
Diet	Fruits: Eats fruit 1+ times/day (Yes/No) Veggies: Eats vegetables 1+ times/day (Yes/No)
Alcohol Consumption	HvyAlcoholConsump: Heavy drinking (Yes/No)

determine if the demographic variable differs across diabetes status. Overall, 84.24% of the respondents reported no diabetes, 1.83% were prediabetes and 13.93% had diabetes. More of the respondents were female (56%) than males (44%), and respondents between the age of 35 and 80+ have more than 5% of the total sample in their respective category. Nearly half of the respondents were college graduate (42%) and 36% reported annual household income of $75k or higher. A Pearson Chi-squared test and a one-way Anova was conducted to determine if randomization was balanced between the groups, however all results were significant indicating that the randomization is not balanced and we would have to adjust for covariate in the analysis.

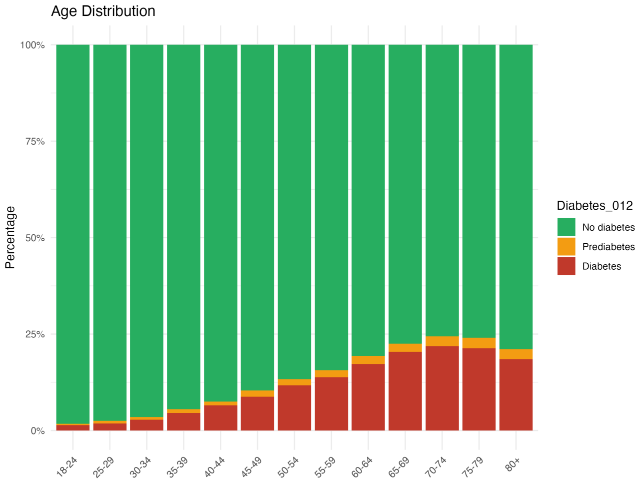


Figure 1. Age Distribution by diabetes status

Figures 1 shows the distribution of diabetes status across age groups and Figure 2 shows the rate of diabetes/prediabetes by age group and sex. The stacked bar chart (Figure 1) shows that the proportion of individuals with diabetes and prediabetes increases steadily with age but reduced at age 75 and above. With figure 2, we observed that both males and females shows an increasing rate of diabetes and prediabetes with age; however, males generally show slightly higher rates across most age categories, particularly between ages 40 and 80+.

Figure 3 below shows the heatmap of the rate of diabetes across education level and household income. The plot showed that lower income and lower household income is associated with high diabetes and prediabetes rate across respondents. We saw that respondents who never attended school or only kindergarten and those with elementary education whose household income is less than or equal to $15,000 show diabetes rates exceeding 40% compared to some college or college graduate who have the same household income. As both education and income level increase, diabetes prevalence decreases consistently, this shows the impact of these two variables on respondents' health outcome.

Table 2 shows the distribution of health behaviors, mental and physical health indicators, cardiovascular comorbidities, and healthcare access across individuals with no diabetes,

TABLE 2. Demographic characteristics across Diabetes Status

Variable	**Overall** (N = 253,680)	**No Diabetes** (N = 213,703; 84.24%)	**Prediabetes** (N = 4,631; 1.83%)	**Diabetes** (N = 35,346; 13.93%)	**p-value**
Sex					
Female	141,974 (56%)	120,959 (57%)	2,604 (56%)	18,411 (52%)	<0.001
Male	111,706 (44%)	92,744 (43%)	2,027 (44%)	16,935 (48%)	
Age					
18–24	5,700 (2.2%)	5,601 (2.6%)	21 (0.5%)	78 (0.2%)	<0.001
25–29	7,598 (3.0%)	7,404 (3.5%)	54 (1.2%)	140 (0.4%)	
30–34	11,123 (4.4%)	10,737 (5.0%)	72 (1.6%)	314 (0.9%)	
35–39	13,823 (5.4%)	13,055 (6.1%)	142 (3.1%)	626 (1.8%)	
40–44	16,157 (6.4%)	14,943 (7.0%)	163 (3.5%)	1,051 (3.0%)	
45–49	19,819 (7.8%)	17,765 (8.3%)	312 (6.7%)	1,742 (4.9%)	
50–54	26,314 (10%)	22,808 (11%)	418 (9.0%)	3,088 (8.7%)	
55–59	30,832 (12%)	26,019 (12%)	550 (12%)	4,263 (12%)	
60–64	33,244 (13%)	26,809 (13%)	702 (15%)	5,733 (16%)	
65–69	32,194 (13%)	24,939 (12%)	697 (15%)	6,558 (19%)	
70–74	23,533 (9.3%)	17,790 (8.3%)	602 (13%)	5,141 (15%)	
75–79	15,980 (6.3%)	12,132 (5.7%)	445 (9.6%)	3,403 (9.6%)	
80+	17,363 (6.8%)	13,701 (6.4%)	453 (9.8%)	3,209 (9.1%)	
Education					
Never attended school	174 (<0.1%)	125 (<0.1%)	2 (<0.1%)	47 (0.1%)	<0.001
Elementary	4,043 (1.6%)	2,699 (1.3%)	161 (3.5%)	1,183 (3.3%)	
Some high school	9,478 (3.7%)	6,868 (3.2%)	314 (6.8%)	2,296 (6.5%)	
High school graduate	62,750 (25%)	50,334 (24%)	1,350 (29%)	11,066 (31%)	
Some college	69,910 (28%)	58,223 (27%)	1,333 (29%)	10,354 (29%)	
College graduate	107,325 (42%)	95,454 (45%)	1,471 (32%)	10,400 (29%)	
Income					
<$10k	9,811 (3.9%)	7,114 (3.3%)	314 (6.8%)	2,383 (6.7%)	<0.001
$10k–$15k	11,783 (4.6%)	8,341 (3.9%)	356 (7.7%)	3,086 (8.7%)	
$15k–$20k	15,994 (6.3%)	12,005 (5.6%)	421 (9.1%)	3,568 (10%)	
$20k–$25k	20,135 (7.9%)	15,622 (7.3%)	459 (9.9%)	4,054 (11%)	
$25k–$35k	25,883 (10%)	20,792 (9.7%)	587 (13%)	4,504 (13%)	
$35k–$50k	36,470 (14%)	30,431 (14%)	748 (16%)	5,291 (15%)	
$50k–$75k	43,219 (17%)	37,219 (17%)	735 (16%)	5,265 (15%)	
$75k+	90,385 (36%)	82,179 (38%)	1,011 (22%)	7,195 (20%)	

n (%); Pearson's Chi-squared test for categorical variables.

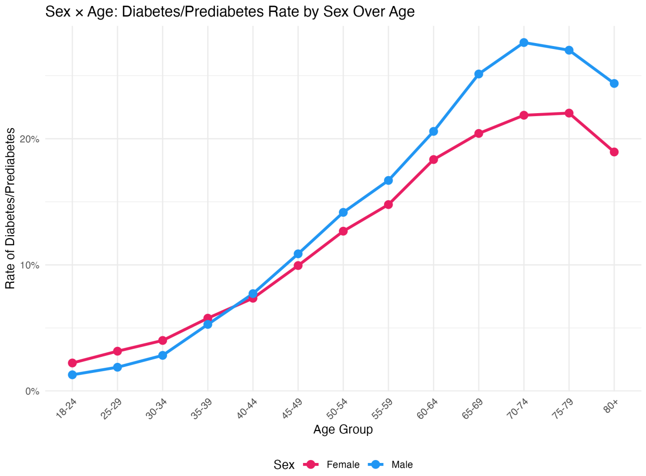


Figure 2. Rate of diabetes/prediabetes by sex over age

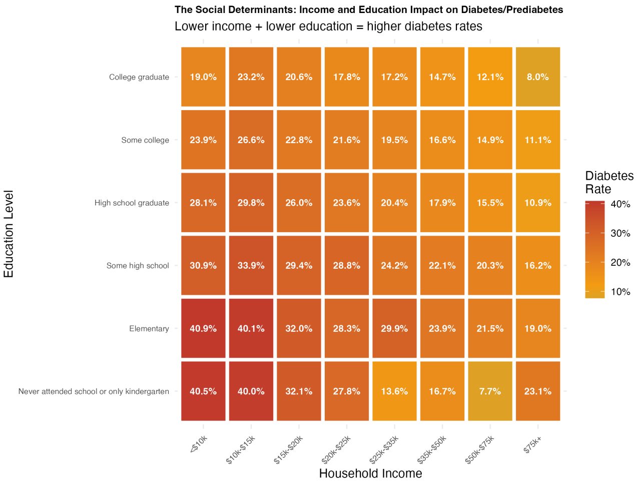


Figure 3. Relationship between education level, household income, and diabetes prevalence.

prediabetes, and diabetes. Categorical variables were compared using Pearson's Chi-squared test, while continuous variables were analyzed using a one-way ANOVA to evaluate differences across diabetes status groups.

Overall, 43% of respondents reported having high blood pressure, and 42% had high cholesterol, but these two conditions were prevalent among those with diabetes and prediabetes.

The mean BMI increased across groups, from 28 among non-diabetic participants to 32 among those with diabetes.

Figures 3 and 4 shows the compounding effects of multiple metabolic and lifestyle risk factors specifically high blood pressure, high cholesterol, heart disease, smoking, and obesity (BMI > 30) on diabetes prevalence, as well as the moderating role of physical activity across different risk levels.

The number of risk factor was obtained by counting the number of times all the five conditions (high blood pressure, high cholesterol, heart disease, smoking, and obesity (BMI > 30)) appear in each of the diabetes status level. As shown in the plot, the prevalence of diabetes increases as the number of risk factor present increases. For participants with no identified risk factors, only about 2% have diabetes and 1% are prediabetes, 97% do not have diabetes. We observed that the proportion of those who have diabetes increase to 52% for participants who has all the 5 number of risk present.

For Figure 4, the metabolic risk level was gotten using the following.

- Very High Risk: Respondents with both high blood pressure, high cholesterol, and a BMI greater than 30
- High Risk: Respondents with either high blood pressure or high cholesterol, and a BMI greater than 30.
- Moderate Risk: Respondents with high blood pressure, or high cholesterol, or BMI > 30. Low Risk: Respondents with none of these risk indicators.

The result showed that exercise physical activity decreases the rate of diabetes across different risk level. This shows exercise could be helpful in reducing the risk of diabetes across respondents.

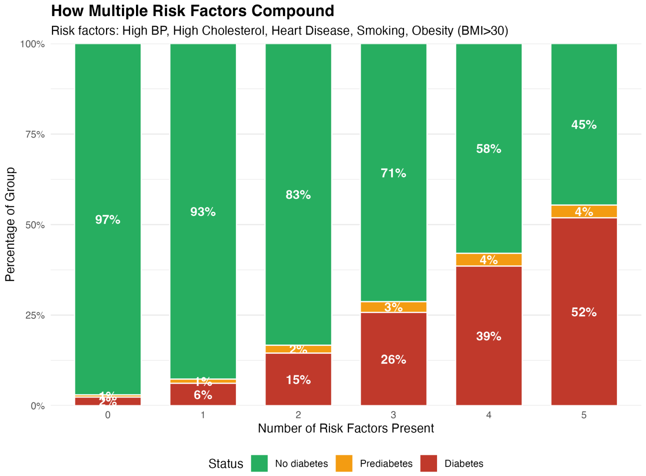


Figure 4. Percentage showing the number of risk factor present for each diabetes status

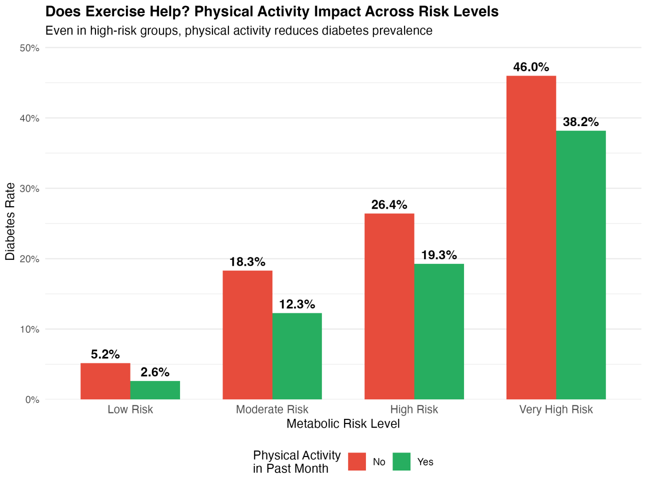
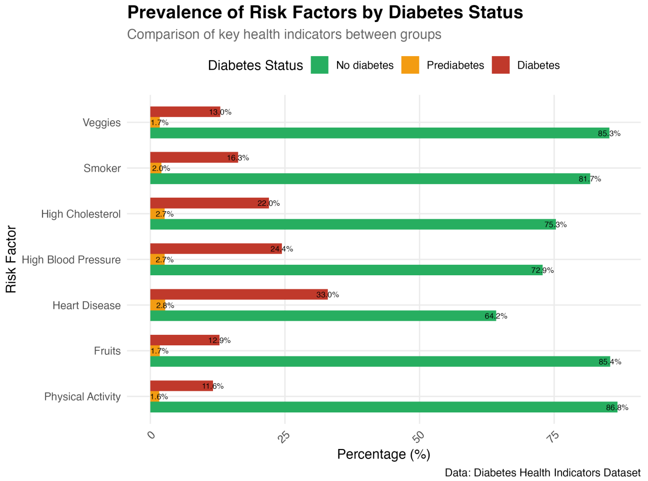


Figure 5. How physical activity and metabolic risk level on Diabetes Prevalence

Figure 6 below shows the proportion of risk factors across diabetes status. The plot shows an increase in the proportion of respondents who has diabetes for heart disease, high blood pressure and high cholesterol. However, respondents who does physical activity, eat fruits and veggies, smokes have lower proportion of diabetes.



Data: Diabetes Health Indicators Dataset

Figure 6. Prevalence of risk factor by diabetes status

Statistical inferences

To understand the various predictors of diabetes status, this project used two statistical methods: Logistic regression and Baseline category Logit model. The response variable initially had three categories: No diabetes (n = 213,703), Prediabetes (4,631) and Diabetes (35,346). For the binary logistic model, both the prediabetes and diabetes levels were combined to form a single level named "Diabetes", which result in a binary response variable (No diabetes = 0, Diabetes = 1). The Baseline category Logit model however retained the three categories to fit the model.

Multiple Logistic Regression

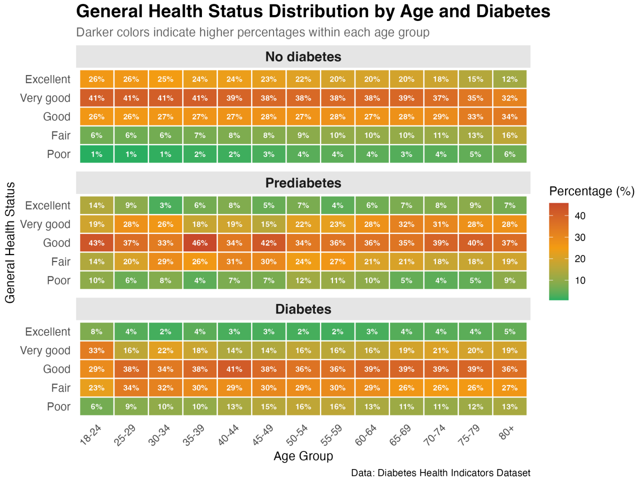
TABLE 3. Descriptive table for health behaviors, mental and physical health status, cardiovascular comorbidities, and healthcare access across Diabetes Status

Variable	**Overall**	**No Diabetes**	**Prediabetes**	**Diabetes**	**p-value**
	(N = 253,680)	(N = 213,703)	(N = 4,631)	(N = 35,346)	
HighBP					
No high BP	144,851 (57%)	134,391 (63%)	1,718 (37%)	8,742 (25%)	<0.001
High BP	108,829 (43%)	79,312 (37%)	2,913 (63%)	26,604 (75%)	
HighChol					
No high cholesterol	146,089 (58%)	132,673 (62%)	1,756 (38%)	11,660 (33%)	<0.001
High cholesterol	107,591 (42%)	81,030 (38%)	2,875 (62%)	23,686 (67%)	
CholCheck					
No check in 5 years	9,470 (3.7%)	9,167 (4.3%)	62 (1.3%)	241 (0.7%)	<0.001
Checked in 5 years	244,210 (96%)	204,536 (96%)	4,569 (99%)	35,105 (99%)	
BMI (Mean ± SD)	28 (±7)	28 (±6)	31 (±7)	32 (±7)	<0.001
Smoker					
Yes	112,423 (44%)	91,824 (43%)	2,282 (49%)	18,317 (52%)	<0.001
Stroke					
Yes	10,292 (4.1%)	6,759 (3.2%)	265 (5.7%)	3,268 (9.2%)	<0.001
Heart Disease/Attack					
Yes	23,893 (9.4%)	15,351 (7.2%)	664 (14%)	7,878 (22%)	<0.001
PhysActivity					
Yes	191,920 (76%)	166,491 (78%)	3,142 (68%)	22,287 (63%)	<0.001
Fruits					
Yes	160,898 (63%)	137,416 (64%)	2,789 (60%)	20,693 (59%)	<0.001
Veggies					
Yes	205,841 (81%)	175,544 (82%)	3,561 (77%)	26,736 (76%)	<0.001
Heavy Alcohol Consumption					
Yes	14,256 (5.6%)	13,216 (6.2%)	208 (4.5%)	832 (2.4%)	<0.001
Any Healthcare					
Yes	241,263 (95%)	202,962 (95%)	4,377 (95%)	33,924 (96%)	<0.001
No Doctor Because of Cost					
Yes	21,354 (8.4%)	17,013 (8.0%)	599 (13%)	3,742 (11%)	<0.001
General Health					
Poor/Excellent	12,081 (4.8%)	7,152 (3.3%)	351 (7.6%)	4,578 (13%)	<0.001
Fair	31,570 (12%)	20,755 (9.7%)	1,025 (22%)	9,790 (28%)	
Good	75,646 (30%)	60,461 (28%)	1,728 (37%)	13,457 (38%)	
Very Good	89,084 (35%)	81,489 (38%)	1,214 (26%)	6,381 (18%)	
Excellent	45,299 (18%)	43,846 (21%)	313 (6.8%)	1,140 (3.2%)	
Mental Health (Mean ± SD)	3 (±7)	3 (±7)	5 (±9)	4 (±9)	<0.001
Physical Health (Mean ± SD)	4 (±9)	4 (±8)	6 (±10)	8 (±11)	<0.001
Difficulty Walking					
Yes	42,675 (17%)	28,269 (13%)	1,285 (28%)	13,121 (37%)	<0.001

n(%); Mean (±SD). Pearson's Chi-squared test for categorical variables; One-way ANOVA for continuous variables.

The multiple Logistic regression method was used as it models the relationship between a binary outcome variable and multiple explanatory variables (both continuous and categorical). Since we already created a binary outcome variable (No diabetes = 0, Prediabetes/Diabetes = 1), then a multiple logistic regression can be used to estimate the odds of having prediabetes/diabetes with respect to regressors such as Demographic, Body Mass Index, High Blood Pressure, High Cholesterol, Smoking, Health General and Mental Health, Chronic Health Conditions, Health Care, Physical Activity, Diet, and Alcohol Consumption. This project used a purposeful selection of the explanatory variables by including the main effects and variables were retained based on statistical significance, their role as a potential confounder and their importance to the study using a backward elimination stepwise algorithm. The multiple logistic regression has the form:

$$\begin{aligned}\text{Logit}(P(Y=1)) &= \log\left(\frac{P(Y=\text{Prediabetes/Diabetes})}{P(Y=\text{No Diabetes})}\right)\\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k\end{aligned}$$



Data: Diabetes Health Indicators Dataset

Figure 7. General Health status distribution by Age and Diabetes

Baseline Category Logit Model

A baseline category logit model is a generalization of the logistic regression to multiple nominal categories. It assumes the response variable follows a multinomial distribution and since our dataset initially has three categories (diabetes, Prediabetes and Diabetes), we use this model to fit the data based on the same regressor as used with Logistic regression.

This model has c − 1 = 3 − 1 = 2 equations, corresponding to comparisons as shown in the equation below

$$\log\left(\frac{\pi_{\text{Diabetes}}}{\pi_{\text{No Diabetes}}}\right) = \alpha_{\text{D}} + \beta_{1\text{D}}X_1 + \beta_{2\text{D}}X_2 + \cdots + \beta_{k\text{D}}X_k$$

$$\log\left(\frac{\pi_{\text{Prediabetes}}}{\pi_{\text{No Diabetes}}}\right) = \alpha_{\text{P}} + \beta_{1\text{P}}X_1 + \beta_{2\text{P}}X_2 + \cdots + \beta_{k\text{P}}X_k$$

This model was selected to extend the multiple binary logistic models so that we can have the three response categories in the model. This model allows us to separately compare for prediabetes and diabetes.

Model inference, knowledge discovery, and discussion.

What did you learn about the data and experiments?

The dataset contained a large and diverse sample of participants (n 253,680) with information on demographic, lifestyle, and health-related characteristics. The response variable captured three levels of diabetes status which are: No diabetes (213,703), Prediabetes (4,631), and Diabetes (35,346).

Among the regressors, four were ordinal in nature: GenHlth, Age, Education, and Income. To determine the most appropriate way to model these ordered variables, two logistic regression models were compared: 1) treating the above four variables in a quantitative manner, 2) treating them in a categorical manner. The result from both comparison (LRT Statistic: 1340, df = 24, p < 0.001) showed that treating these variables as a factor fit better than treating it in a quantitative manner, thus for other modelling, these regressors were used as a factor.

Relationship between response and predictors.

Logistic Regression

A multiple logistic regression was conducted to determine the association between diabetes status (prediabetes/diabetes vs. no diabetes) and regressors such as regressors such as Demographic, Body Mass Index, High Blood Pressure, High Cholesterol, Smoking, Health General and Mental Health, Chronic Health Conditions, Health Care, Physical Activity, Diet, and Alcohol Consumption. Two models were compared: 1) Main effect only and 2) Main effect with some selected predictors. The result from both comparison (LRT Statistic: 609.35, df = 52, p < 0.001) showed that the model with the interaction fits better and thus was used.

Interaction Effects

Note: Only significant interaction is reported here.

BMI:Age: The effect of BMI on diabetes risk increases with age. Each extra BMI point matters more at older age. A 5-unit BMI increase at age 70 has a stronger diabetes effect than at age 25.

TABLE 4. BMI Effect Across Age Groups

Age Group	Odds Ratio	Interpretation
18–24	$exp(0.0246) = 1.025$	2.5% increase.
40–44	$exp(0.0246 + 0.0247) = 1.0505$	5.05% higher odds.
45–49	$exp(0.0246 + 0.0211) = 1.0468$	4.68% higher odds.
50–54	$exp(0.0246 + 0.0230) = 1.0488$	4.88% higher odds.
55–59	$exp(0.0246 + 0.0271) = 1.0531$	5.31% higher odds.
60–64	$exp(0.0246 + 0.0357) = 1.0622$	6.22% higher odds.
65–69	$exp(0.0246 + 0.0372) = 1.0637$	6.37% higher odds.
70–74	$exp(0.0246 + 0.0430) = 1.070$	7.00% higher odds.
75–79	$exp(0.0246 + 0.0384) = 1.0650$	6.50% higher odds.
80+	$exp(0.0246 + 0.0442) = 1.0712$	7.12% higher odds.

18–24 (baseline); Interpretation: A one-unit increase in BMI multiplies the odds of diabetes by 1.025 holding other predictors constant

Sex × Age: The protective effect of being male reduces with age, with older males having significantly higher odds of diabetes.

General Health × Physical Health: This shows Physical health matter more for people who rate their general health as excellent/very good, and matter less for people already in poor health with regards to the odds of having diabetes.

High Cholesterol × heart disease: People who have both high cholesterol and heart disease have exp(0.610 + 0.466 -0.342) = exp(0.724) = 2.06 times the odds of diabetes

TABLE 5. Sex × Age Interaction Effects on Diabetes Odds

Age Group	Odds Ratio	Interpretation
18–24	$exp(-0.5230) = 0.5927$	40.76% lower odds
40–44	$exp(-0.5230 + 0.5432) = 1.0205$	2.05% higher odds
45–49	$exp(-0.5230 + 0.6056) = 1.0861$	8.61% higher odds.
50–54	$exp(-0.5230 + 0.6515) = 1.1371$	13.71% higher odds.
55–59	$exp(-0.5230 + 0.6774) = 1.1670$	16.70% higher odds.
60–64	$exp(-0.5230 + 0.7090) = 1.2044$	20.44% higher odds.
65–69	$exp(-0.5230 + 0.8597) = 1.4003$	40.03% higher odds.
70–74	$exp(-0.5230 + 0.9683) = 1.5610$	56.10% higher odds.
75–79	$exp(-0.5230 + 0.9615) = 1.5504$	55.04% higher odds.
80+	$exp(-0.5230 + 1.0048) = 1.6190$	61.90% higher odds.

18–24 (baseline); Males have 0.5927 times the odds of diabetes compared to females at baseline age. 40.76% decrease in the odds of diabetes for males aged 18-24. Males aged 40–44 have about 2.04% higher odds of diabetes than females of the same age, controlling for all other variables in the model. Overall, protective effect of being male disappears and reverses with age.

TABLE 6. General Health × Physical Health Interaction on Diabetes Odds

General Health	Odds Ratio	Interpretation
Poor	$exp(-0.0119) = 0.9882$	Decreases odds by 1.2%.
Fair	$exp(-0.0119 + 0.0052) = 0.9933$	Decreases odds by 0.67%.
Good	$exp(-0.0119 + 0.0115) = 0.9996$	Almost no effect on odds.
Very Good	$exp(-0.0119 + 0.0163) = 1.0044$	Increases odds by 0.45%.
Excellent	$exp(-0.0119 + 0.0556) = 1.0447$	Increases odds by 4.47%.

Note: Physical Health = for how many days during the past 30 days was your physical health not good?; At baseline (poor),each additional day of poor physical health decreases diabetes odds by 1.2%. overall, physical health worsens risk more strongly among healthier individuals.

compared to people who have neither high cholesterol nor heart disease, holding all other predictors constant.

BMI × Physical Activity: Each unit increase in BMI increases the odds of diabetes by 3.07% for physically active people (exp(0.0246 + 0.0056) = 1.0307).

Main Effects

Note: Only significant main effect not associated with any significant interaction effect is reported here.

Key Health Predictors

- High blood pressure (β = 0.68, p < .001), and Cholesterol checked in the past 5 years (β = 1.17, p < .001), were strong positive predictors. This shows that the odds of having prediabetes/diabetes for individuals with high blood pressure is 1.97 times that of those who does not have high blood pressure. Also, the odd of having prediabetes/diabetes for individuals who had their Cholesterol checked in the past 5 years is 3.19 times those who did not check in the last 5 years.

Chronic Health Conditions

- History of stroke (Ever told you have a stroke, (β = 0.14, p < .001)) and heart disease or attack (Ever reported having coronary heart disease (CHD) or myocardial infarction (MI), (β = 0.47, p < .001)) were also significant risk factors.

Diet

- The odds of prediabetes/diabetes (β = -0.02, p =0.08) for Individuals who consume fruit 1 or more times per day is 0.98 times those who do not consume fruit 1 or more times per day.
- The odds of prediabetes/diabetes (β = -0.03, p =0.04) for Individuals who consume vegetables one or more times per day is 0.97 times those who do not consume fruit 1 or more times per day.

Health Care

- Those who reported not seeing a doctor because of cost have about 4% higher odds of diabetes (β = 0.04, p =0.08).

General and Mental Health

- Having a serious difficulty walking or climbing stairs is positively associated with diabetes (β = 0.12, p < .001).
- The number of days during the past 30 days for which individual mental health was not good showed a non-significant negative association with the likelihood of diabetes, (β = -0.001, p =.13).

Demographics

- The odds of having prediabetes/diabetes for individuals with Income $75k+ ($\beta$ = -0.42, p < .001), $50k-$75k (β = -0.28, p < .001), $35k-$50k (β = -0.24, p < .001), $25k-$35k (β = -0.16, p < .001), $20k-$25k (β = -0.09, p = .004), $15k-$20k (β = -0.06, p = .06), $10k-$15k (β = -0.018, p = .59), is 0.66, 0.76, 0.79, 0.85, 0.91, 0.94, 0.98 times those with income <$10k respectively. This means that people with household income of $75k+ have 34% lower odds of having diabetes compared to those with income less than $10k.

Baseline Category Logit Model

A baseline category logit model was conducted to determine the association between diabetes, prediabetes vs. no diabetes. A separate model was created for the response diabetes vs. no diabetes, and prediabetes vs. diabetes. Two models were compared: 1) Main effect only and 2) Main effect with some selected predictors. The result from both comparison (LRT Statistic: 676.73, df = 108, p < 0.001) showed that the model with the interaction fits better and thus was used.

Interaction Effects

Note: Only significant interaction is reported here.

BMI:Age: The effect of BMI on diabetes risk increases with age.

TABLE 7. Age–BMI Interaction Effects on Odds of Diabetes

Age Group	Odds Ratio	Interpretation
Diabetes		
18–24	$\exp(0.0223) = 1.0226$	2.26% higher odds.
35–39	$\exp(0.0223 + 0.0269) = 1.0504$	5.04% higher odds.
40–44	$\exp(0.0223 + 0.0287) = 1.0523$	5.23% higher odds.
50–54	$\exp(0.0223 + 0.0265) = 1.0500$	5.00% higher odds.
55–59	$\exp(0.0223 + 0.0304) = 1.0541$	5.41% higher odds.
60–64	$\exp(0.0223 + 0.0389) = 1.0631$	6.31% higher odds.
65–69	$\exp(0.0223 + 0.0410) = 1.0653$	6.53% higher odds.
70–74	$\exp(0.0223 + 0.0473) = 1.0721$	7.21% higher odds.
75–79	$\exp(0.0223 + 0.0412) = 1.0656$	6.56% higher odds.
80+	$\exp(0.0223 + 0.0491) = 1.0740$	7.40% higher odds.
Prediabetes		
No significant interaction between age and BMI for prediabetes.		

Baseline = 18–24; For diabetes, a one-unit increase in BMI multiplies the odds of diabetes by 1.0226 holding other predictors constant. 2.3% odds

Sex × Age: The protective effect of being male reduces with age, with older males having significantly higher odds of diabetes. Same trend was observed for prediabetes.

TABLE 8. Sex–Age Interaction Effects on Odds of Diabetes and Prediabetes

Age Group	Odds Ratio	Interpretation
Diabetes		
18–24	$\exp(-0.3593) = 0.6982$	Lower odd at baseline age.
45–49	$\exp(-0.3593+0.4867) = 1.1359$	13.59% higher odds.
50–54	$\exp(-0.3593+0.5029) = 1.1544$	15.44% higher odds.
55–59	$\exp(-0.3593+0.5475) = 1.2071$	20.71% higher odds.
60–64	$\exp(-0.3593+0.5593) = 1.2214$	22.14% higher odds.
65–69	$\exp(-0.3593+0.7123) = 1.4233$	42.33% higher odds.
70–74	$\exp(-0.3593+0.8252) = 1.5934$	59.34% higher odds.
75–79	$\exp(-0.3593+0.8083) = 1.5667$	56.67% higher odds.
80+	$\exp(-0.3593+0.8530) = 1.6384$	63.84% higher odds.
Prediabetes		
18–24	$\exp(-1.2052) = 0.2996$	70.04% lower odds.
35–39	$\exp(-1.2052+1.1744) = 0.9697$	0.3% lower odds.
40–44	$\exp(-1.2052+1.0927) = 0.8936$	10.64% lower odds.
45–49	$\exp(-1.2052+1.0537) = 0.8594$	14.06% lower odds.
50–54	$\exp(-1.2052+1.2373) = 1.0326$	3.26% higher odds.
55–59	$\exp(-1.2052+1.1224) = 0.9205$	7.95% lower odds.
60–64	$\exp(-1.2052+1.2929) = 1.0917$	9.17% higher odds.
65–69	$\exp(-1.2052+1.4169) = 1.2378$	23.78% higher odds.
70–74	$\exp(-1.2052+1.5109) = 1.3576$	35.76% higher odds.
75–79	$\exp(-1.2052+1.5865) = 1.4642$	46.42% higher odds.
80+	$\exp(-1.2052+1.6189) = 1.5124$	51.24% higher odds.

Baseline = 18–24; At baseline for diabetes, Males have 0.6982 times the odds of diabetes compared to females at baseline age.

General Health × Physical Health: This shows Physical health matter more for people who rate their general health as excellent/very good, and matter less for people already in poor health with regards to the odds of having diabetes. Same was also observed for prediabetes.

TABLE 9. Effect of Physical Health Days Across General Health Categories

General Health	Odds Ratio	Interpretation
Diabetes		
Poor	$\exp(-0.0124) = 0.9877$	decreases odds by 1.23%.
Fair	$\exp(-0.0124+0.0054) = 0.9930$	almost no effect
Good	$\exp(-0.0124+0.0112) = 0.9988$	almost no effect
Very Good	$\exp(-0.0124+0.0182) = 1.0058$	increases odds by 0.58%.
Excellent	$\exp(-0.0124+0.0590) = 1.0477$	increases odds by 4.77%.
Prediabetes		
Poor	$\exp(-0.0063) = 0.9937$	decreases odds by 1.18%.
Excellent	$\exp(-0.0063+0.0383) = 1.0325$	increases odds by 3.25%.

Note: Physical health = number of days (0–30) during the past 30 days that physical health was not good; Baseline = poor

Heavy Alcohol Consumption X Age:

TABLE 10. Interaction Between Heavy Alcohol Consumption and Age

Age	Odds Ratio	Interpretation
Diabetes		
No significant interaction between alcohol consumption and age.		
Prediabetes		
18–24.	$\exp(-0.7902) = 0.4538$	Heavy alcohol consumption is associated with lower odds of prediabetes.
30–34	$\exp(-0.7902+1.2649) = 1.6075$	Heavy drinkers have 1.61 times the odds of prediabetes compared with non-heavy drinkers at this age.

High Cholesterol × heart disease (Diabetes): People who have both high cholesterol and heart disease have exp(0.6155 + 0.5025 -0.3583) = 2.14 times the odds of diabetes compared to people who have neither high cholesterol nor heart disease, holding all other predictors constant.

High Cholesterol × heart disease (Pre-Diabetes): People who have both high cholesterol and heart disease have exp(0.5886 + 0.1565 -0.2454) = 1.65 times the odds of pre-diabetes compared to people who have neither high cholesterol nor heart disease, holding all other predictors constant.

BMI × Physical Activity: Each unit increase in BMI increases the odds of diabetes by 2.87% for physically active people (exp(0.0223 + 0.0060) = 1.0287).

Main Effects

The main effect is summarized in *Table 11*

TABLE 11. Significant Main Effects Not Involved in Interactions

Variable	**Diabetes**	**Prediabetes**
High blood pressure	$\beta = 0.7345,\ p < .001$ The odds of diabetes for individuals with high blood pressure is 2.08 times that of those without high blood pressure.	$\beta = 0.3720,\ p < .001$ The odds of prediabetes for individuals with high blood pressure is 1.45 times that of those without high blood pressure.
Cholesterol checked in the past 5 years	$\beta = 1.2514,\ p < .001$ The odds of diabetes for individuals who checked their cholesterol in the past 5 years is 3.50 times those who did not.	$\beta = 0.8221,\ p < .001$ The odds of prediabetes for individuals who checked their cholesterol in the past 5 years is 2.28 times those who did not.
History of stroke	$\beta = 0.1576,\ p < .001$ The odds of diabetes for individuals with stroke history is 1.17 times those with no history.	
Could not see a doctor due to cost		$\beta = 0.3347,\ p < .001$ The odds of prediabetes for individuals unable to see a doctor due to cost is 1.40 times those who were able to see a doctor.
Serious difficulty walking/climbing	$\beta = 0.1370,\ p < .001$ The odds of diabetes for individuals with difficulty walking is 1.14 times those without difficulty.	

Note: Only significant main effects not associated with any significant interaction are reported.

Which variable(s) are more related to the response?

As listed in the previous section, the plots (*figure 8* and *figure 9*) show the most significant predictors for both the logistic regression and the baseline category logit model.

Compare the predictive modeling methods you used in the study.

Table 12 shows the comparisons of all the model that was fitted and comparison was done using AIC.

What are the pros and cons of these methods?

The logistic regression model has several advantages: it achieves a lower AIC, indicating a better fit, uses fewer parameters and is therefore more powerful, and provides simpler interpretations. However, a key limitation is that it reduces the outcome to a binary classification, causing specific information about the prediabetes stage to be lost.

In contrast, the baseline category logit model captures the prediabetes stage and offers a more comprehensive analysis. Its drawbacks include a higher AIC, meaning a worse fit, the use of more parameters (making it less powerful), and a more complex interpretation.

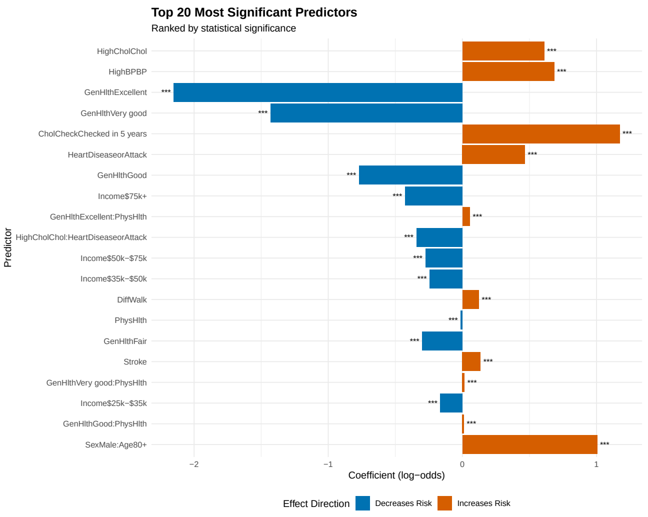


Figure 8. Logistic Model

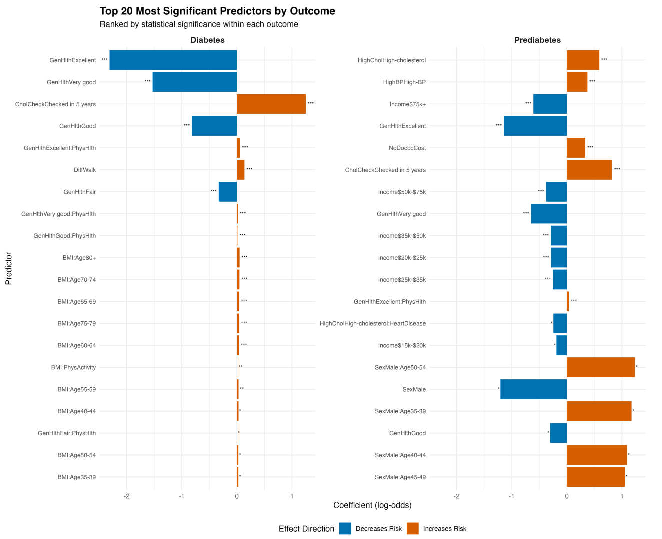


Figure 9. Baseline Category Logit Model

TABLE 12. Model Comparison

Model	Explanatory Variables	Deviance	df	AIC
1	None	221031	253679	
2	ME	174158	253634	174250
3	**ME + I**	**173549**	**253582**	**173745**
4	ME	201805	507268	202013.9
5	ME + I	201224	507206	201553.1

Note: Model 1 = Null Model; Models 2–3 = Logistic Regression; Models 4–5 = Baseline Category Logit Model.

Main Effects (ME): HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, NoDocbcCost, GenHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income.
Interactions (I): BMI:Age, GenHlth:PhysHlth, HighChol:HeartDiseaseorAttack, Sex:Age, BMI:PhysActivity, HvyAlcoholConsump:Age, Smoker:Age.

Is there any interesting discovery from your analysis?

Several interesting discoveries was found in the result. Based on the model with the best fit,

- The effect of diabetes/prediabetes risk became more pronounced in older age groups. At age 18 to 24 years, increase in BMI only increases the risk of diabetes by 2.5% but this increases to 7.1% for those aged 80 years and above suggesting that older people must pay more and closer attention to their weight to reduce their risk of diabetes. Weight management is important in diabetes prevention among older adults.
- There was a reversal in the effect of gender on the risk of diabetes as age increases. At age 18-24, males have about 41% lower odds of diabetes compared to females; however, this effect diminishes as age increases. This could suggest possible shift in body composition or hormonal changes, or how they pay attention to their health.
- Also, as the number of days during the past 30 days for which the respondent's physical health not good increases, diabetes/prediabetes risk only increases among those with very good and excellent general health while the risk reduces with those who have poor or fair general health. This suggests that physical health reduction has more effect on the risk of diabetes for people who good and excellent general health compared to those whose health are already bad.
- Having both high cholesterol and heart diseases doubles the odds of diabetes.
- BMI has a stronger effect on diabetes risk among physical active people compared to inactive people which is counterintuitive.

Conclusions

This project analyze how demographic characteristics, health behaviors, mental and physical health status, cardiovascular comorbidities and healthcare access contributes to the risk of diabetes and prediabetes using the Centers for Disease Control and Prevention 2015 Behavioral Risk Factor Surveillance System survey dataset using two modelling approach (Logistic Regression and Baseline category logit model). The result of both models was compared using AIC with Logistics regression performing better than the baseline category logit model even though the baseline model models take into consideration the three categories of the response (prediabetes, diabetes and no diabetes).

Overall, the Logistic model shows the influences of both demographics and other clinical predictors of diabetes/prediabetes. The influence of BMI on diabetes/prediabetes increases with age, suggesting that more focus should be placed on adults with higher weights. Patients with combination of high cholesterol and heart disease should be paid keen attention as both effect doubles the odds of diabetes/prediabetes. We also saw that, differences in diabetes/prediabetes reverses with age suggesting changes in physical, behavioral or health lifestyle which could further be invested in the next phase of the analysis. We also saw that the risk of diabetes/prediabetes has more influence on those whose general health status is good but has an increased number of days for which their physical health was not good.

Together, this project was able to select the best model and determine the factors that could affect prediabetes and diabetes risk among people.

References

1. Gao, Z. (2025). Trends in diabetes prevalence and key influencing factors in the United States (1950-2021). Advances in Economics, Management and Political Sciences, 166, 175-182.
2. Centers for Disease Control and Prevention. (2024, May 15). National Diabetes Statistics Report. U.S. Department of Health and Human Services. https://www.cdc.gov/diabetes/php/data-research/index.html
3. Winer, N., & Sowers, J. R. (2013). Epidemiology of diabetes. Journal of Clinical Pharmacology, 44(4), 397-405. https://doi.org/10.1177/0091270004263017