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

Categorical Data Analysis Project (BSDS 6210)

Oluwafunmibi Omotayo Fasanya

Louisiana State University Health Science Center, New Orleans.

November 17, 2025

Outline

Introduction

- Diabetes is a chronic condition in which the body either doesn't produce enough insulin or can't use it effectively, causing glucose to build up in the bloodstream instead of being used for energy.
- According to CDC, diabetes is among the top 10 leading causes of death in the U.S.:
  - About 38.4 million Americans (11.6% of the U.S. population) currently have diabetes.
  - 29.7 million have been diagnosed, while 8.7 million (22.8%) are unaware they have the condition.
  - About 97.6 million U.S. adults (38% of the population) have prediabetes, a condition that increases the risk of developing Type 2 diabetes.
- **Goal:** To analyze how demographic characteristics, health behaviors, mental and physical health status, cardiovascular comorbidities, and healthcare access contribute to the risk of diabetes and prediabetes using the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset.

Data Description

- The Diabetes Dataset used was obtained from Kaggle, originally from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS).
- It is a cross-sectional telephone survey of over 400,000 U.S. adults with data colected on health behaviors, chronic conditions and preventive service use.
- The data used contains 253,680 individual and 21 variables.
  - Target variable: Diabetes_012
    - 0 = No diabetes or only during pregnancy
    - 1 = Prediabetes
    - 2 = Diabetes
  - Predictor variables:
    - **Demographic:** Sex, Age, Education, Income
    - **Health:** BMI, Blood Pressure, High cholesterol, Cholesterol check in the past 5 years, General Health, Physical/Mental Health, Difficulty walking or climbing stairs.
    - **Lifestyle:** Smoking, Alcohol, Physical Activity, Diet (Fruits, Vegetables)
    - **Chronic Conditions:** Stroke, Heart Disease
    - **Healthcare Access:** Have any health care coverage or insurance?, Skipped doctor visit due to cost (past year)?

Data Cleaning and Preparation

- The dataset was pre-cleaned, with no missing values.
- Additional data processing performed include:
  - Factor encoding: Converted categorical variables into factors and assigned descriptive labels to their levels.
  - Ordinal encoding: Ordered variables such as GenHlth, Age, Education, and Income.
  - Binary response creation:
    - Overall, 84.24% of the respondents reported no diabetes, 1.83% were prediabetes and 13.93% had diabetes.
    - Combined Prediabetes and Diabetes into one category to form a binary outcome variable
    - 0 = No diabetes or only during pregnancy; 1 = Prediabetes or Diabetes
- Final dataset retained all 21 explanatory variables for modeling.

Exploratory Data Analysis

Demographic and Health Characteristics

Exploratory Data Analysis: Demographic Summary

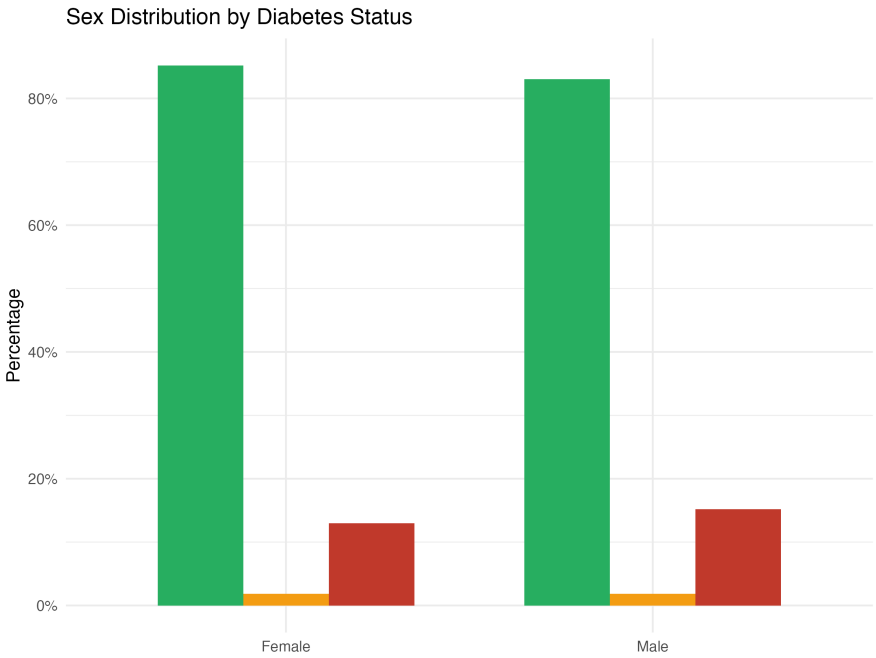

Figure 1a: Sex Distribution by diabetes status

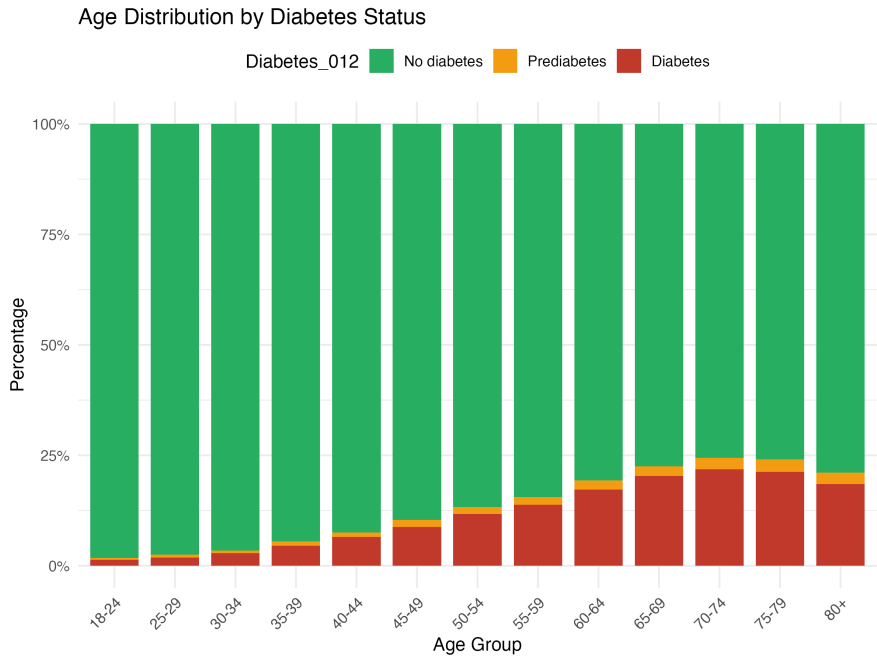

Figure 1b: Age Distribution by diabetes

- Both male and female have almost similar distribution of status.
- Proportion of individuals with diabetes and prediabetes increases steadily with age but reduced at age 75 and above.

Exploratory Data Analysis: Demographic Summary

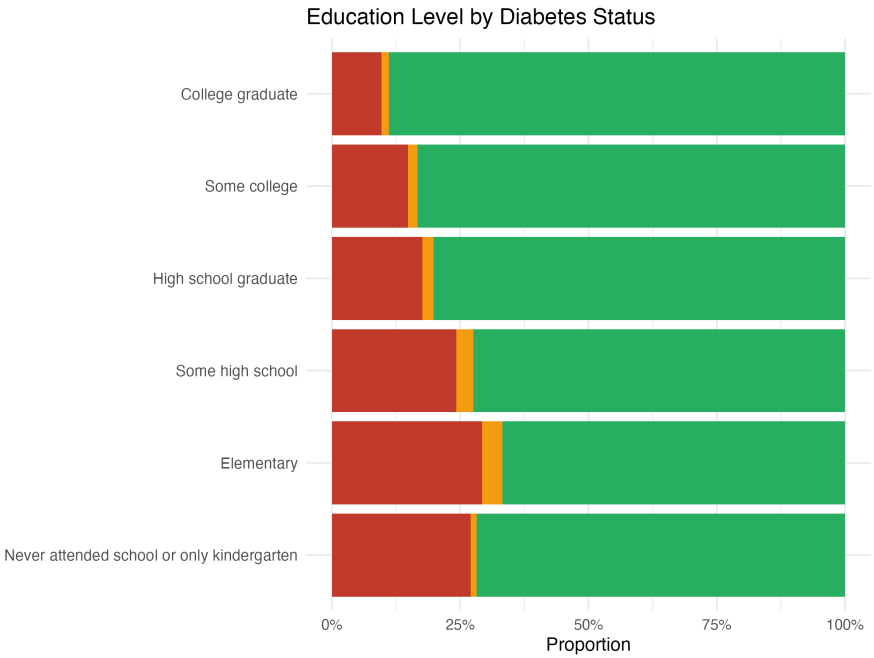

Figure 2a: Education Distribution by diabetes status

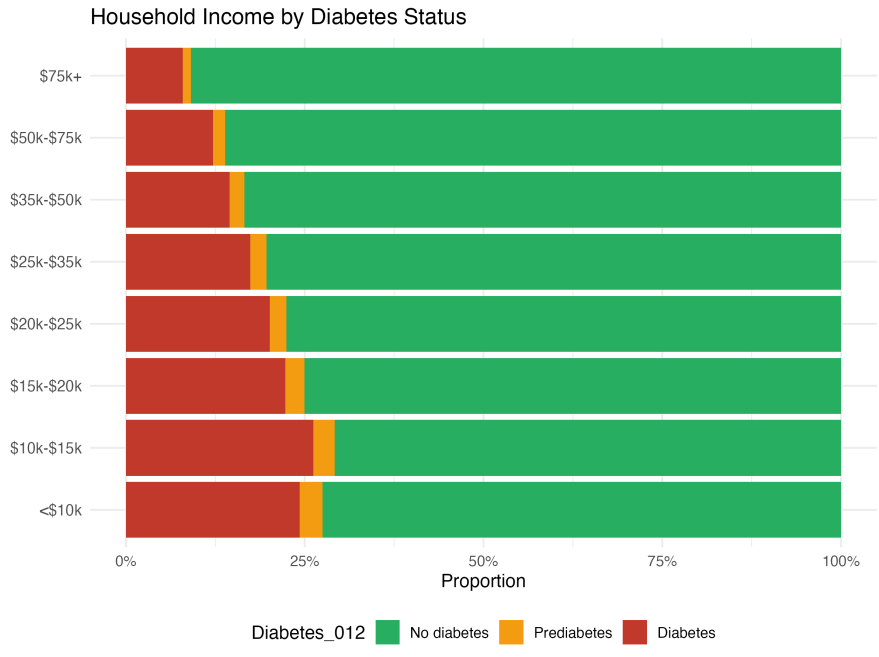

Figure 2b: Income Distribution by diabetes status

- Higher education level and household income shows lower proportion of prediabetes and diabetes

Exploratory Data Analysis: Demographic Summary

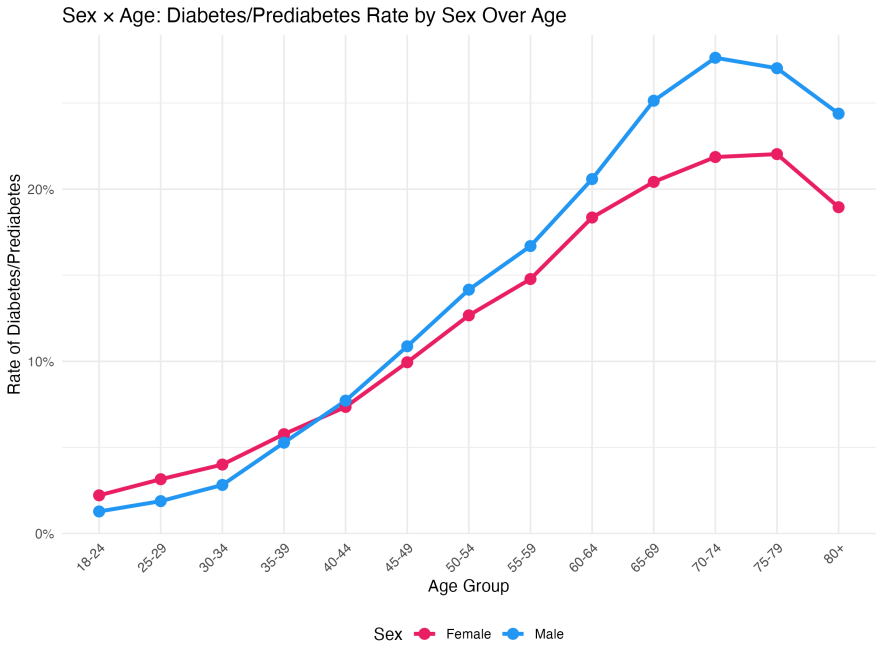

Figure 3a: Sex over Age

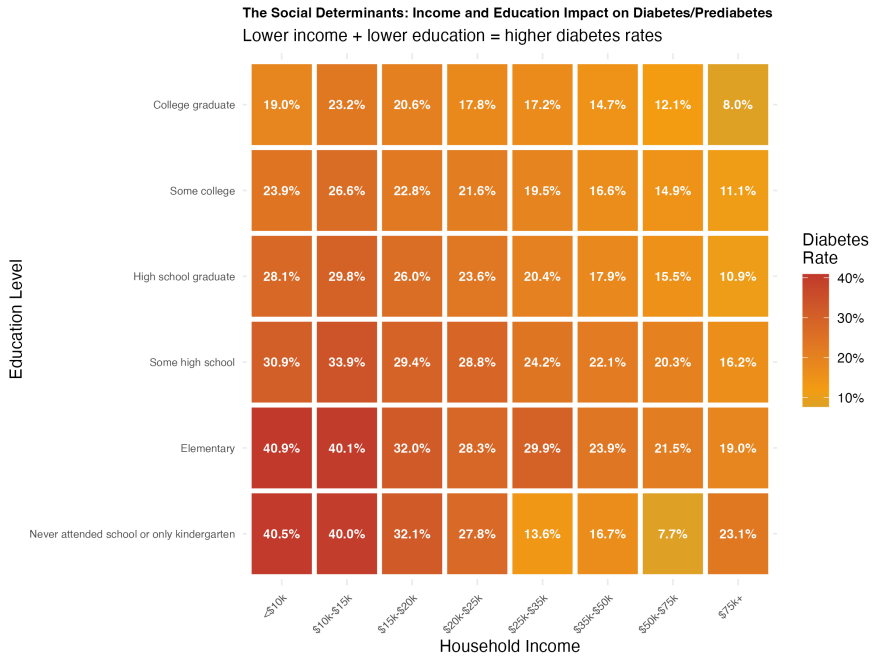

Figure 3b: Education and household income

- Both males and females shows an increasing rate of diabetes and prediabetes with age with a drop at 75-79
- Males generally show slightly higher rates across most age categories, particularly between ages 40 and 80 above
- As education and income level increase, diabetes prevalence decreases.

Exploratory Data Analysis: Health Characteristics

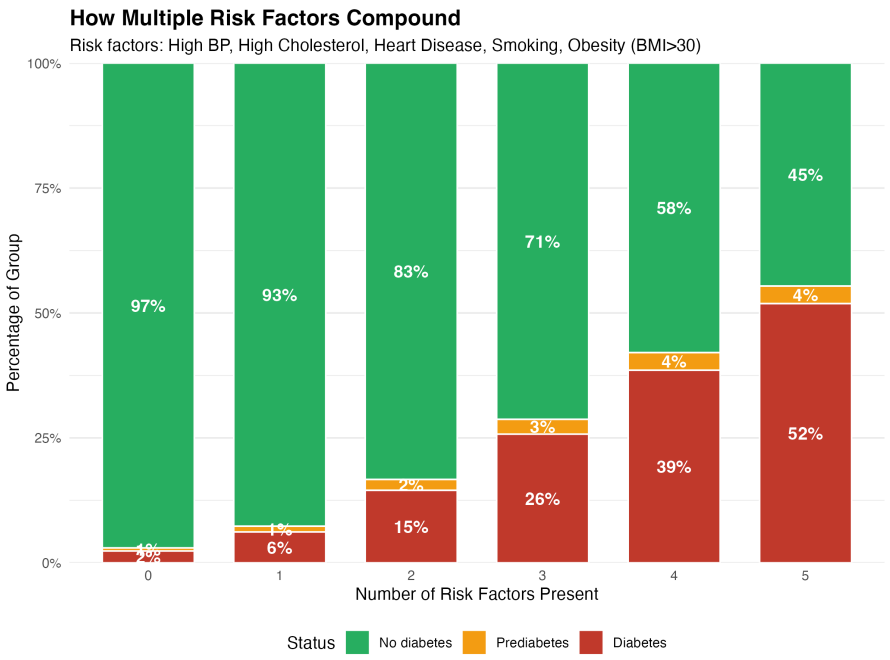

Figure 4a : Number of risk factor present

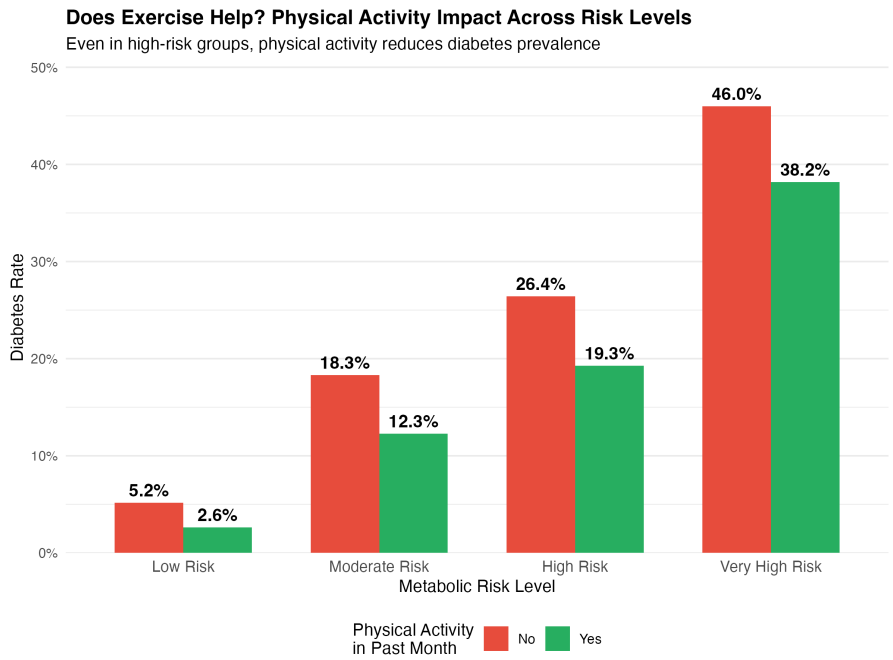

Figure 4b: Physical activity and metabolic

- Prevalence of diabetes increases as the number of risk factor present increases.
- Physical activity decreases the rate of diabetes across different risk level.

Exploratory Data Analysis: Health Characteristics

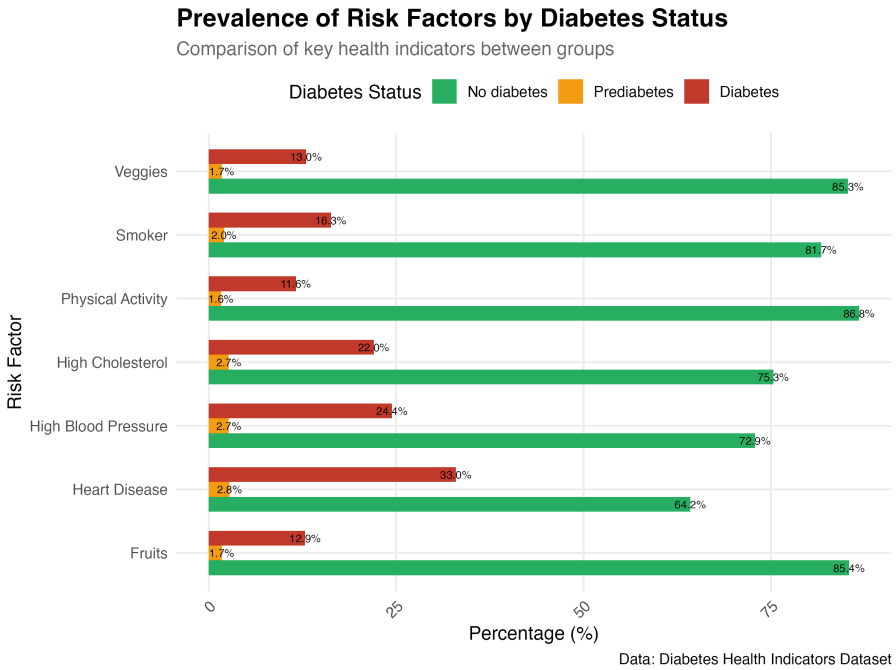

Figure 5a : Prevalence of risk factor

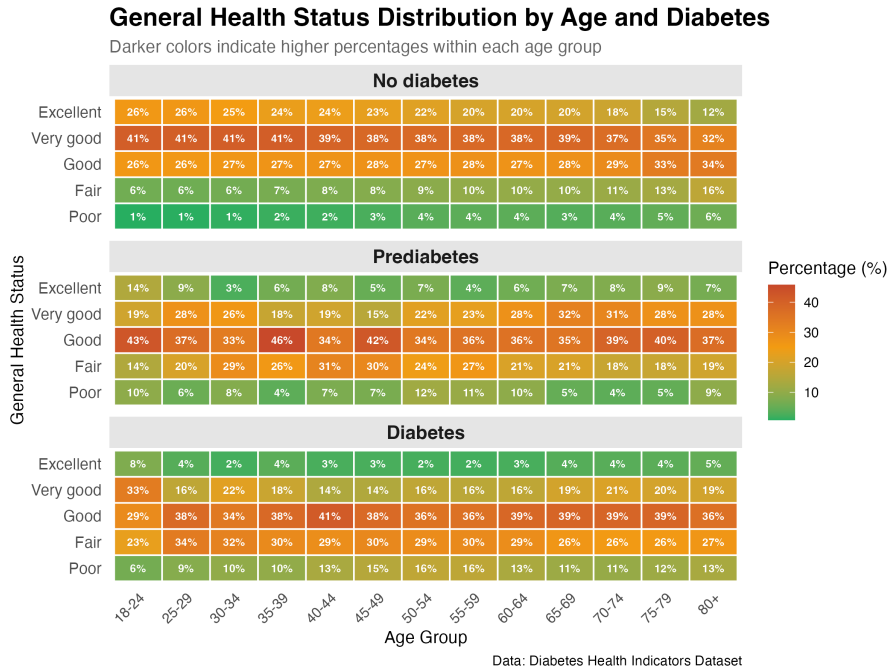

Figure 5b: General Health status distribution

- Heart disease, high blood pressure and high cholesterol is associated with higher proportion of diabetes and prediabetes.
- Physical activity, eat fruits and veggies, smokes have lower proportion of diabetes and prediabetes.

Statistical Inference

Logistic Regression and Baseline Category Logit Model

Statistical Inference: Overview and Logistic Regression

- To identify predictors of diabetes status, two statistical models were used:
  - Multiple Logistic Regression
  - Baseline Category Logit Model
- The **multiple logistic regression** models the log odds of having Diabetes or Prediabetes as a function of explanatory variables:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

- A **purposeful selection** approach with backward elimination was used, retaining significant variables and confounders.

Statistical Inference: Baseline Category Logit Model

- The **Baseline Category Logit Model** generalizes logistic regression for multi-category outcomes.
- Assumes the response follows a multinomial distribution.
- Retains all three outcome categories: *No Diabetes*, *Prediabetes*, and *Diabetes*.
- The model estimates two equations:

$$\log\left(\frac{\pi_{\text{Diabetes}}}{\pi_{\text{No Diabetes}}}\right) = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1k}X_k$$

$$\log\left(\frac{\pi_{\text{Prediabetes}}}{\pi_{\text{No Diabetes}}}\right) = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2k}X_k$$

- Enables separate estimation of predictors influencing each diabetes stage.

Model Inference, Knowledge Discovery, and Discussion

Logistic Regression and Baseline Category Logit Model

What We Learned About the Data and Experiments

- Dataset included **253,680 respondents** with demographic, lifestyle, and health-related characteristics.
- Diabetes status had three categories:
  - No diabetes (213,703; 84.24%)
  - Prediabetes (4,631; 1.83%)
  - Diabetes (35,346; 13.93%)
- Four variables were ordinal: *General Health, Age, Education, and Income*.
- Compared modeling these as:
  1. Quantitative
  2. Categorical (factor)
- Likelihood Ratio Test ($LRT = 1340$, $df = 24$, $p < 0.001$) indicated treating them as **categorical (factor)** gave a significantly better model fit

Multiple Logistic Regression: Main Effects

- **Higher risk:** High blood pressure, stroke, difficulty walking, and checking cholesterol within 5 years.
- **Lower risk:** Higher income levels, regular physical activity, and frequent vegetable consumption.
- *Note: Significant main effects associated with significant interaction effects are not interpreted here.*

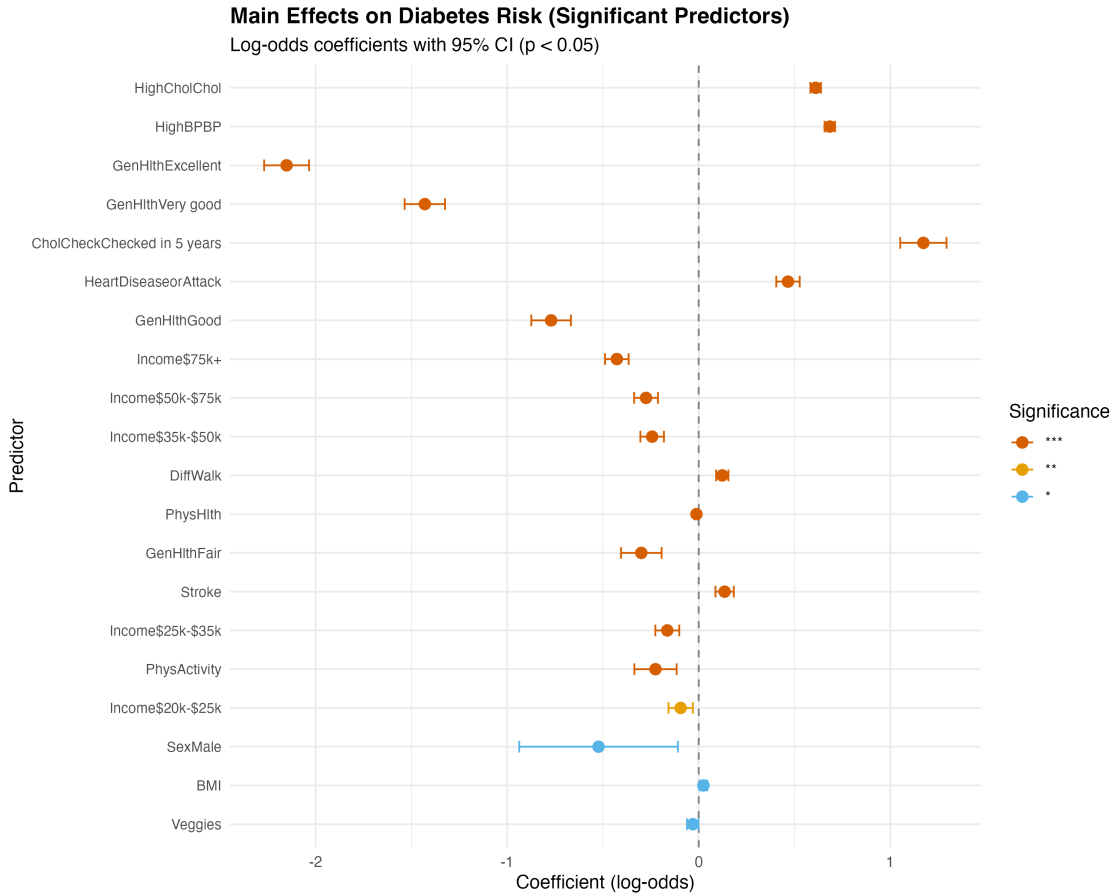

Figure 6a: Main Effects on Diabetes Risk

Relationship Between Response and Predictors

Multiple Logistic Regression: Main Effects

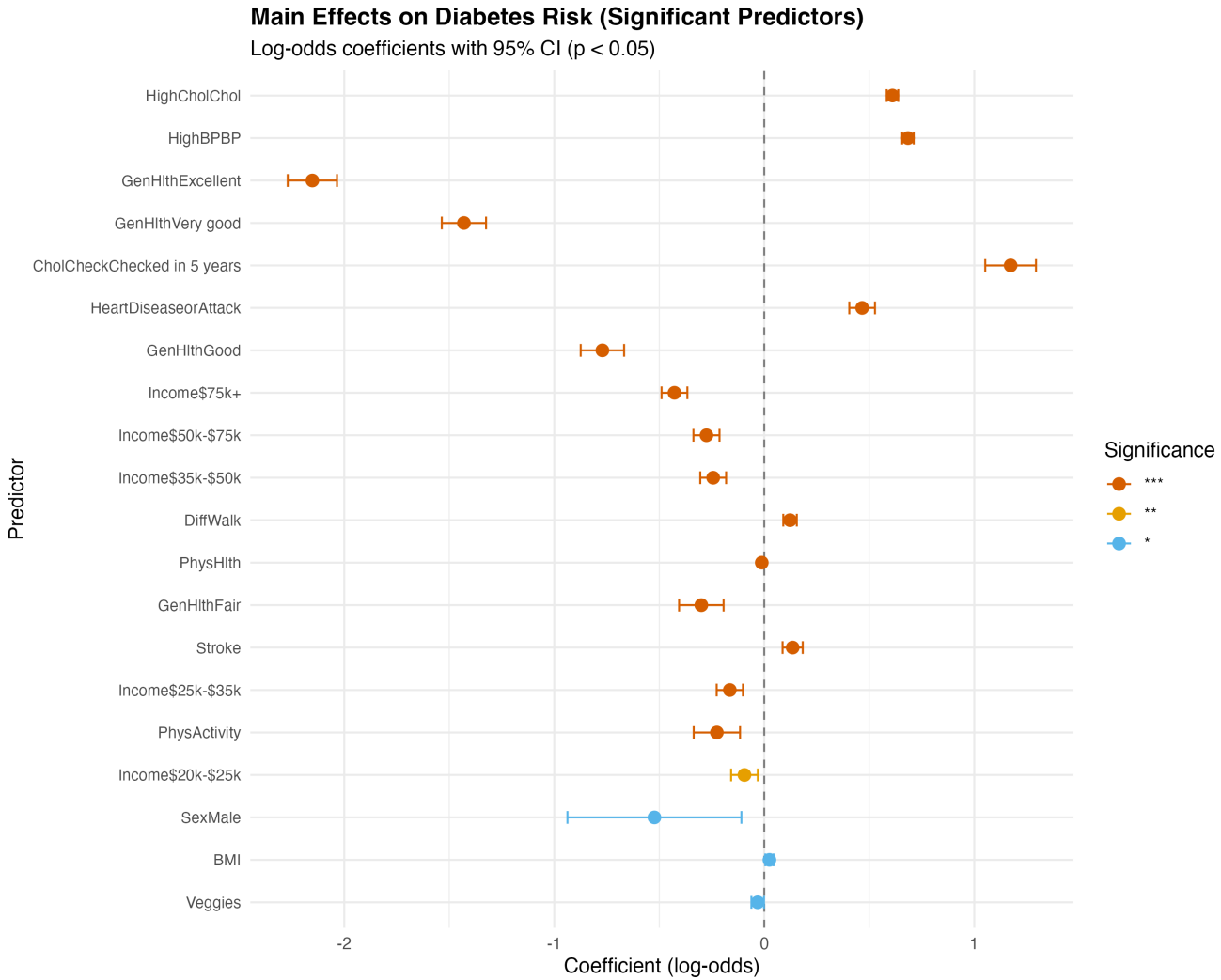

Figure 6b: Main Effects on Diabetes Risk

Multiple Logistic Regression: Interactions

- BMI effect increases with age (2.5% odds increase per BMI unit in young → 71% in older adults)
- Males have lower risk at age 18–24 (47.71% lower) but risk increase increases as age increase (up to 62% higher) than females
- For excellent(4.47%) and very good (0.45%) general health, each additional day of poor physical health increases diabetes odds.
- Having both high cholesterol and heart disease doubles the odds of diabetes compared to having neither condition.
- BMI increases diabetes risk slightly more in physically active individuals (3.1% per BMI unit) than inactive individuals.

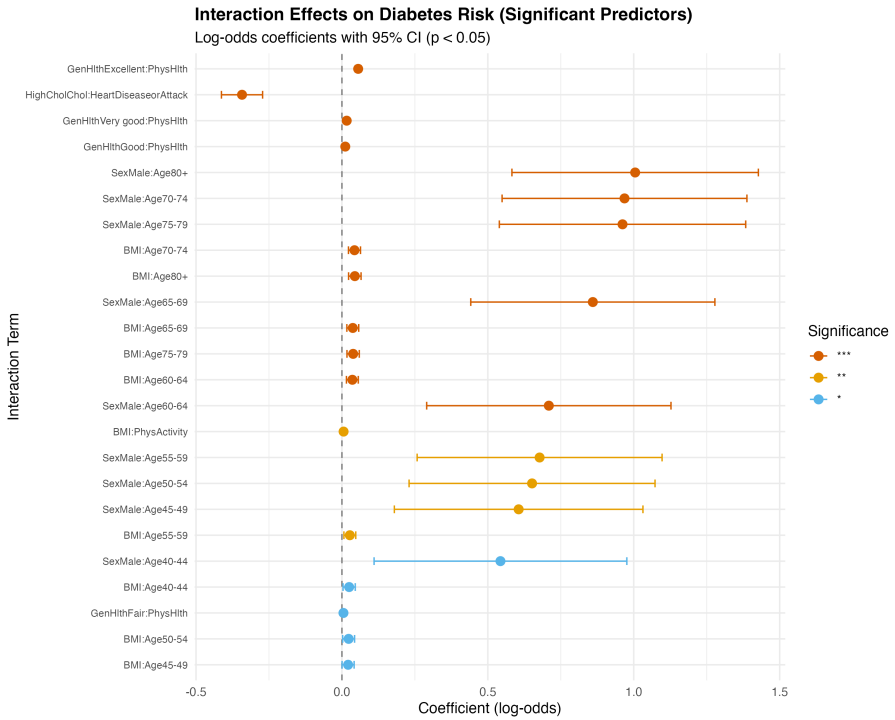

Figure 6b: Interaction Effects on Diabetes

Relationship Between Response and Predictors

Multiple Logistic Regression: Interactions

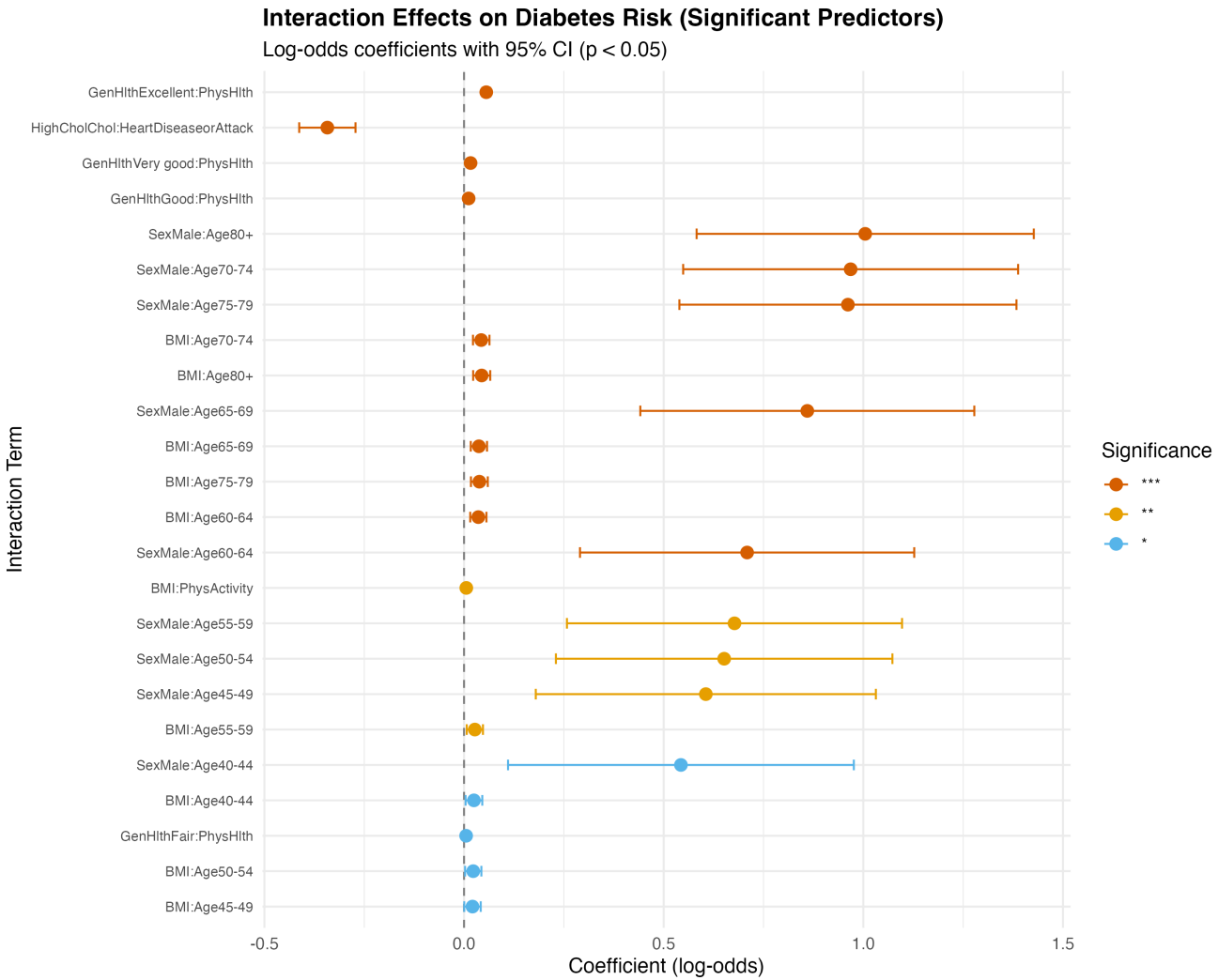

Figure 6b: Interaction Effects on Diabetes

Baseline Category Logit Model: Main Effects

Diabetes Patterns:

- **Risk factors:** Cholesterol check in past 5 years, High blood pressure, High cholesterol, Heart disease/attack, Difficulty walking, Stroke
- **Protective:** Excellent, Very good, and Good general health, Income, Physical activity

Prediabetes Patterns:

- **Risk factors:** Cholesterol check, High BP, High cholesterol, Missed doctor visit in past year due to cost
- **Protective:** Excellent health, Very good health, Good health, Higher income, Physical activity

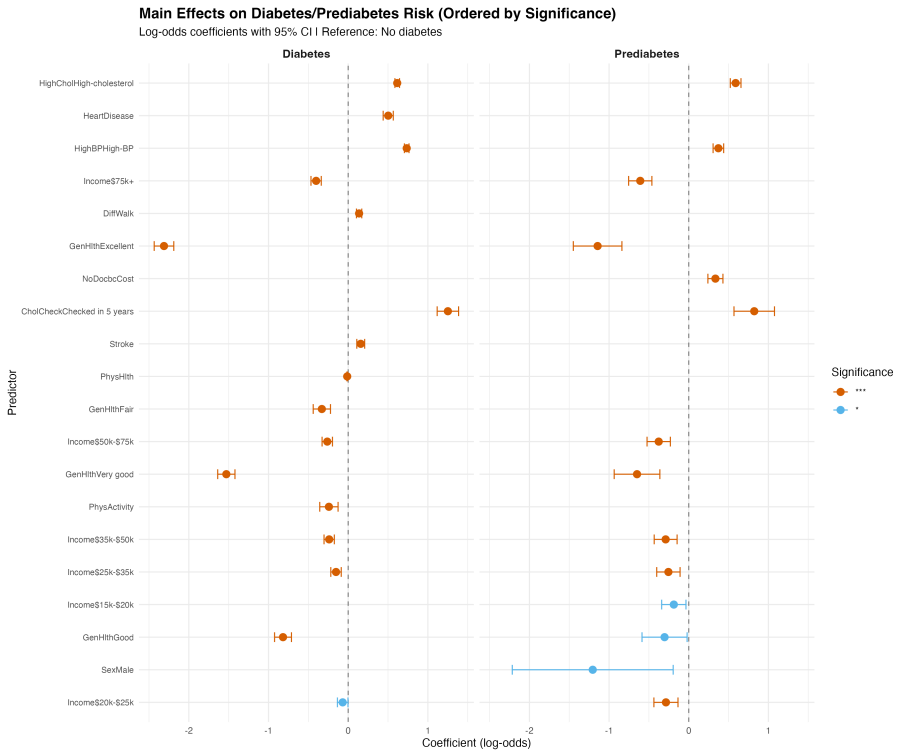

Figure 7a: Main Effects

Relationship Between Response and Predictors

Baseline Category Logit Model: Main Effects

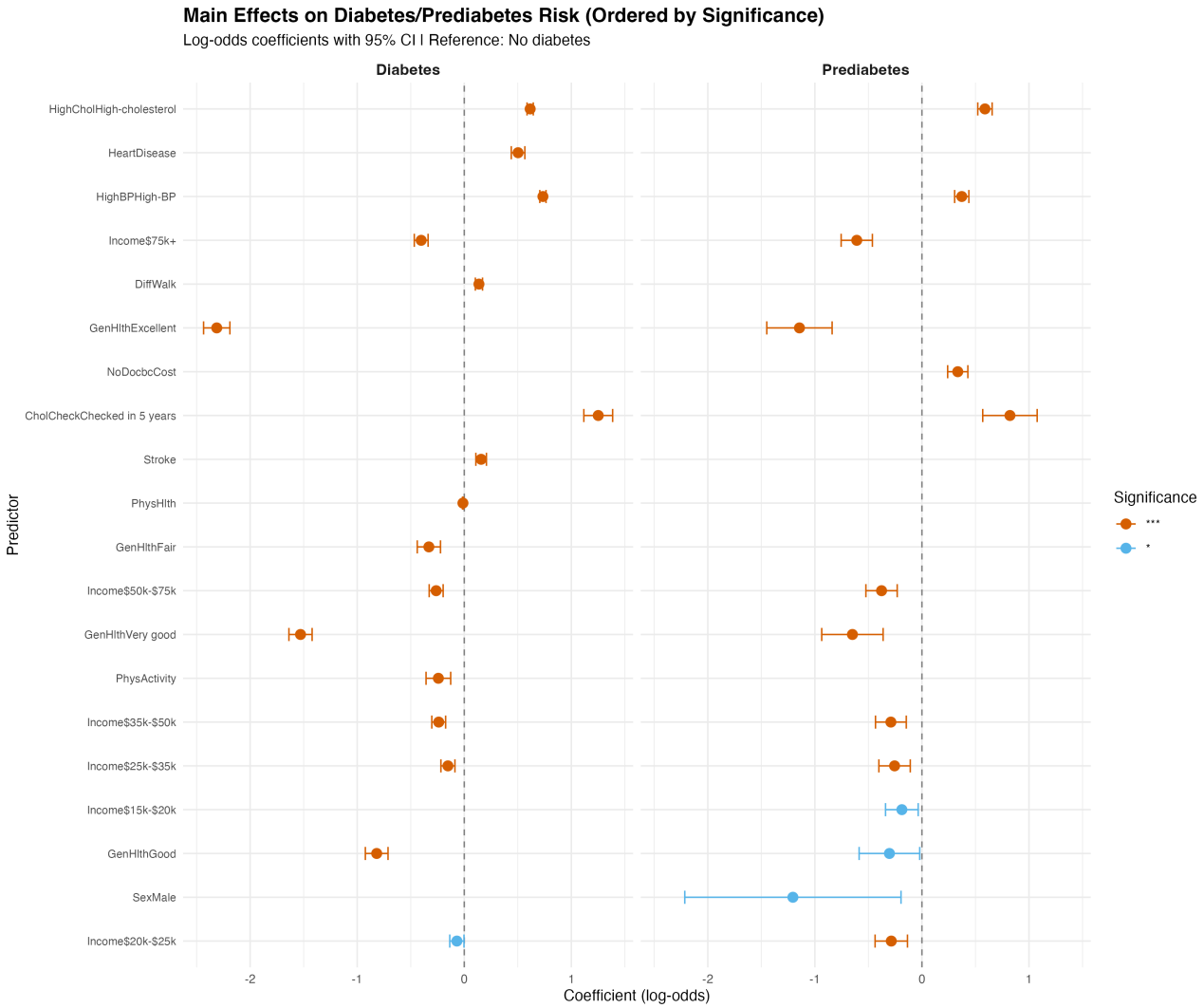

Figure 7a: Main Effects on Diabetes Risk

Baseline Category Logit Model: Interactions

Diabetes Interactions:

- **Sex** $\times$ **Age:** Older males face higher odds than older females.
- **BMI** $\times$ **Age:** BMI effects increases with age.
- **Heavy Alcohol Consumption** $\times$ **Age:** Respondents age 30-34 who are heavy alcohol consumer have a significant higher odd of diabetes.
- **High Cholesterol** $\times$ **Heart Disease:** Having both high cholesterol and heart disease or attack increases odds of diabetes.

Prediabetes Interactions:

- **Sex** $\times$ **Age:** Older males face higher odds than older females.
- **High Cholesterol** $\times$ **Heart Disease:** Having both high cholesterol and heart disease or attack increases odds of prediabetes.

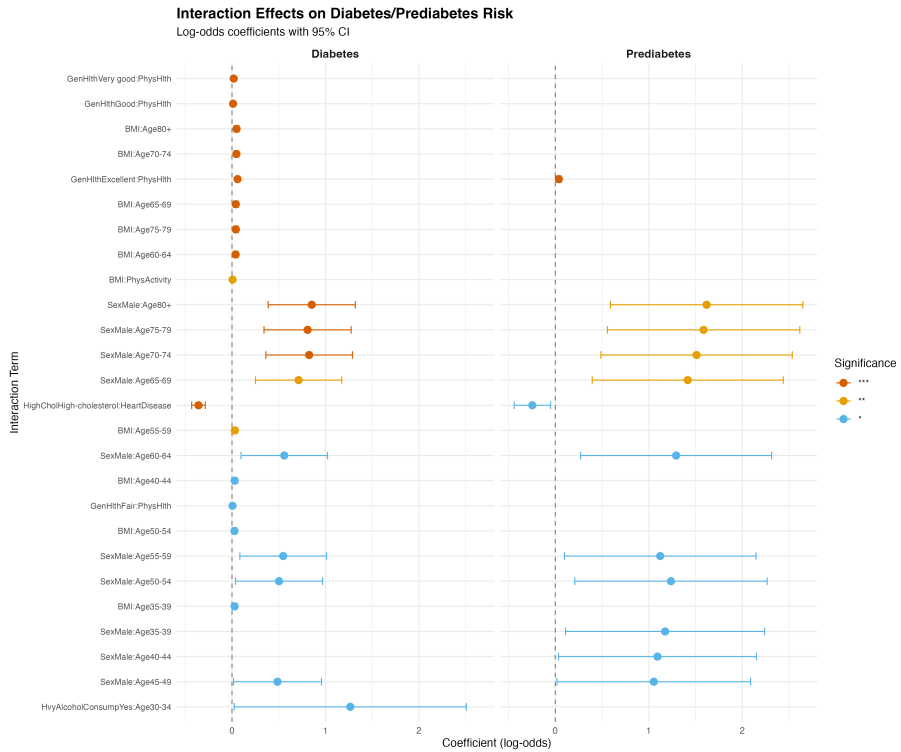

Figure 7b: Interaction Effects Ordered by Significance

Baseline Category Logit Model: Interactions

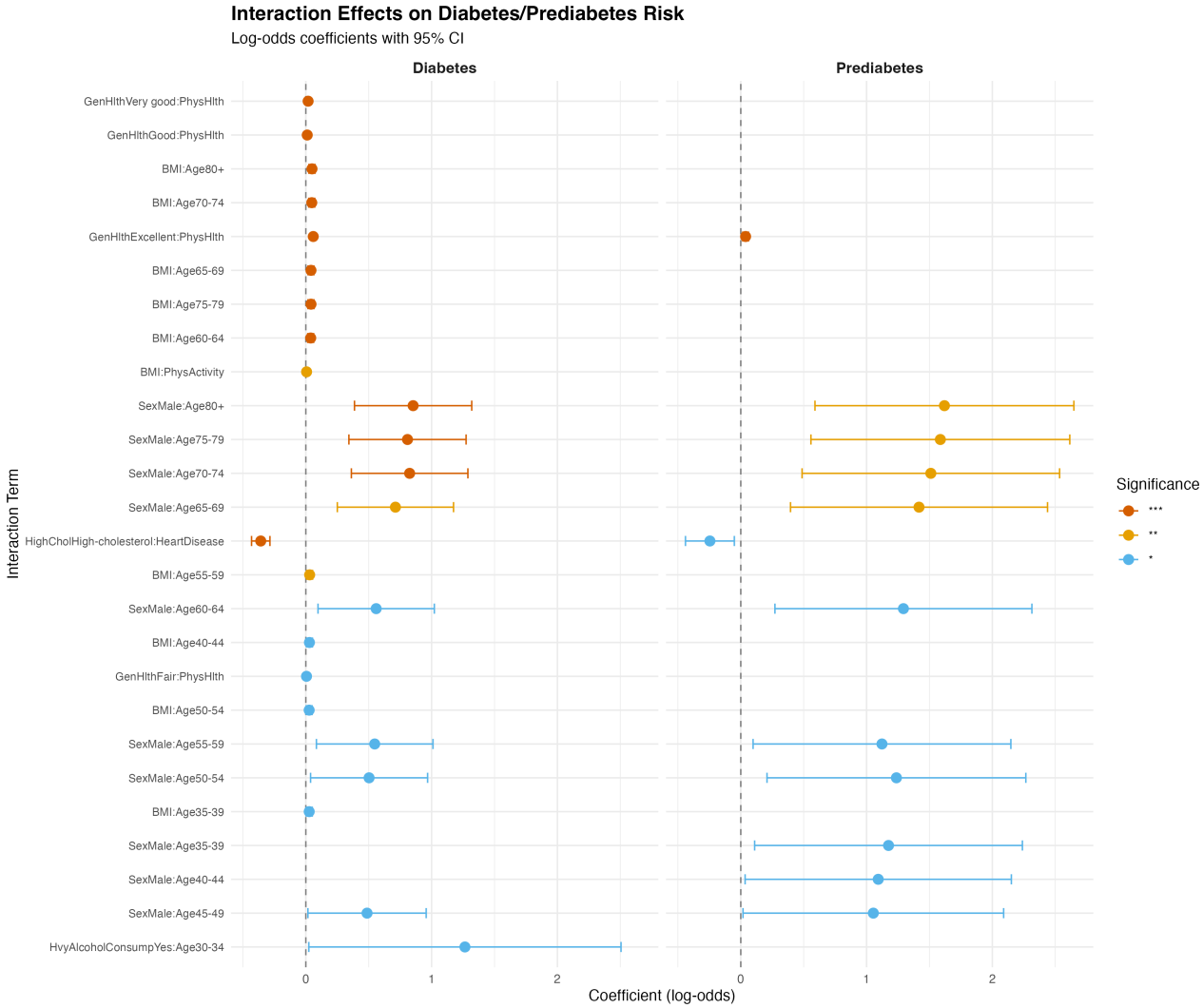

Figure 7b: Interaction Effects Ordered by Significance

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

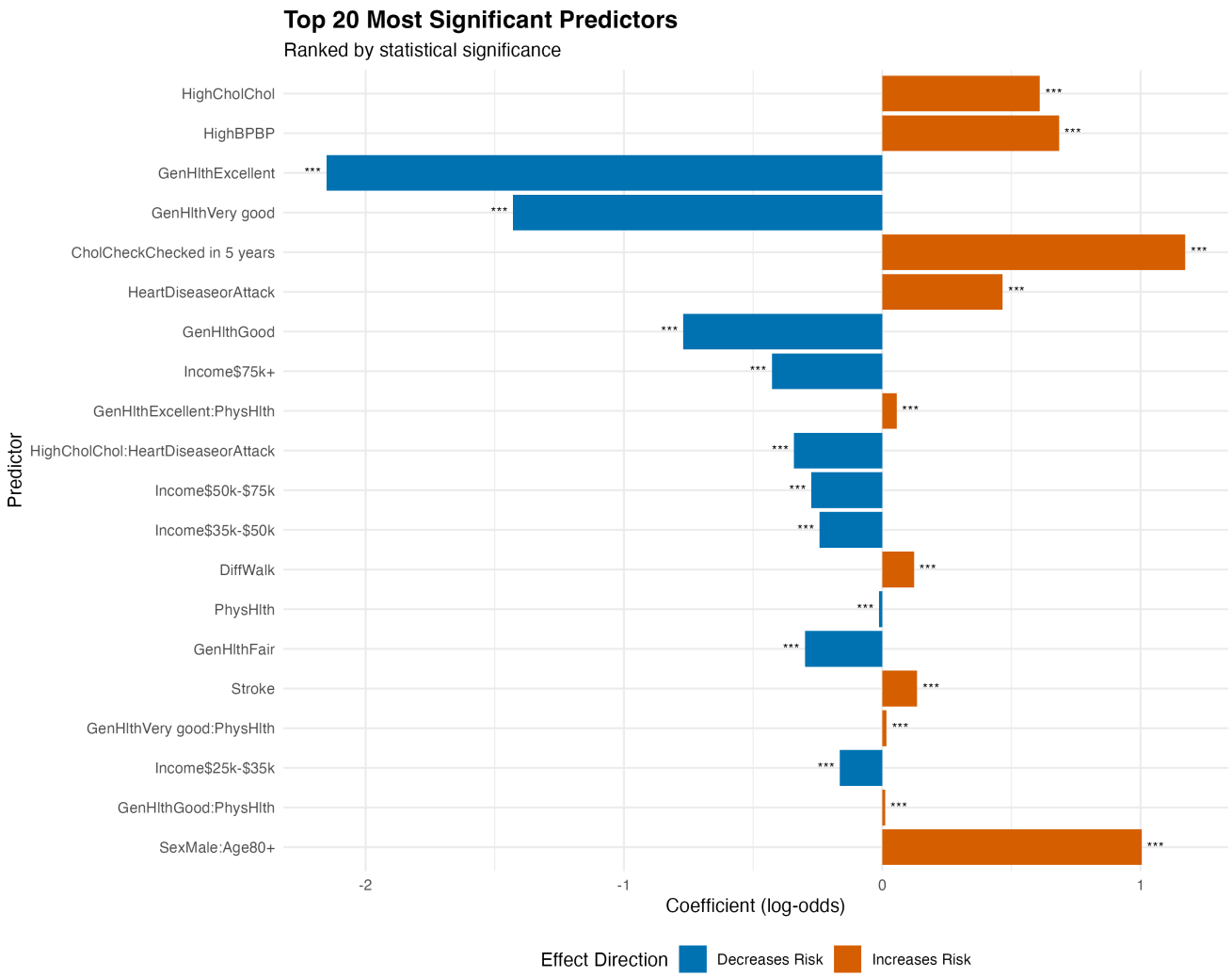

Figure 8a : Logistics Model

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

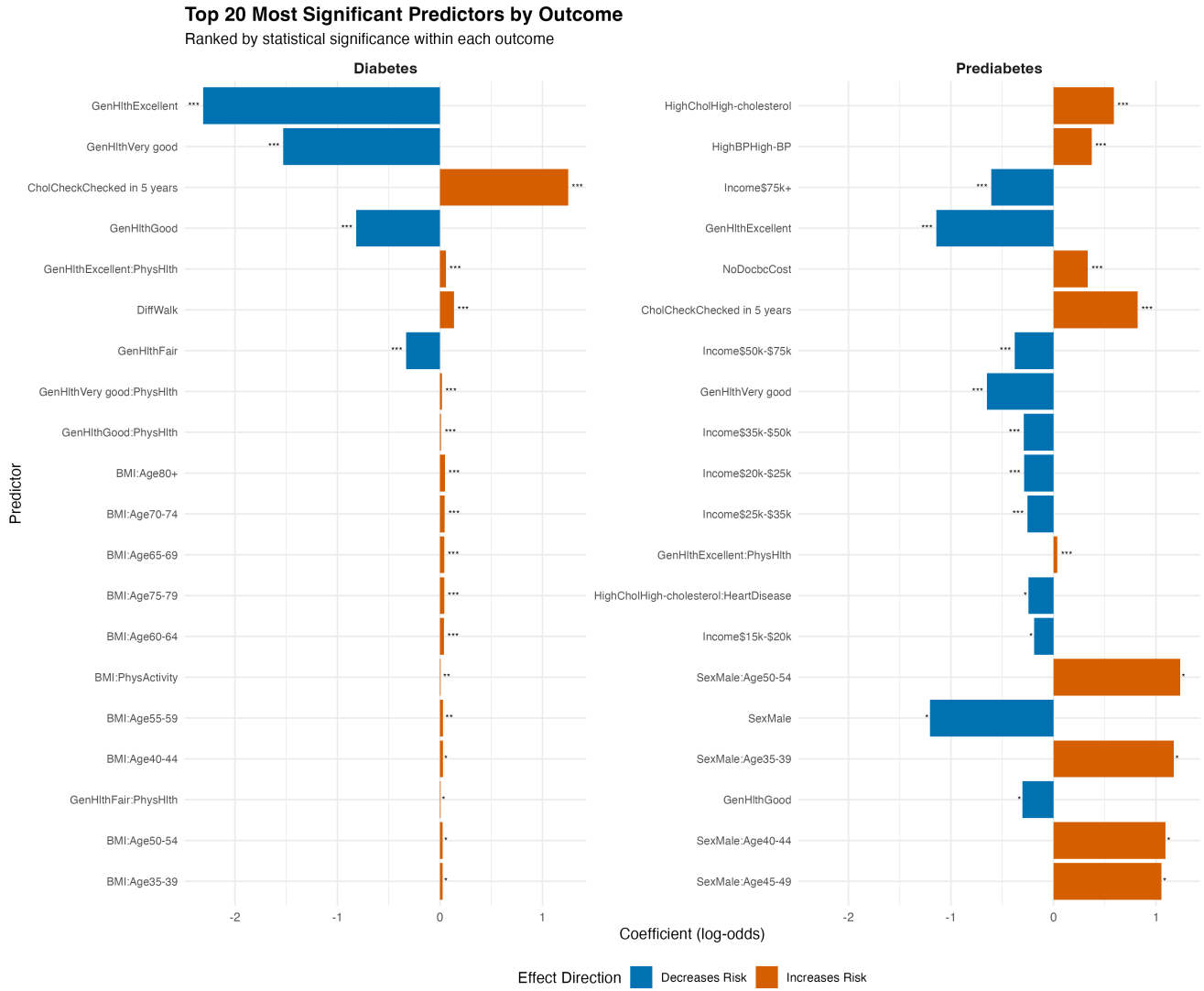

Figure 8b: Baseline Category Logit Model

Model Comparison Summary: Pros and Cons

Model	Explanatory Variables	Deviance	df	AIC
1	None	221031	253679	—
2	Main Effects (ME)	174158	253634	174250
3	ME + Interactions (I)	173549	253582	173745
4	Main Effects (ME)	201805	507268	201989
5	ME + Interactions (I)	201224	507206	201553.1

Notes:

- 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

Binary Logistic (Models 2 & 3):

+ Lower AIC (better fit)
+ Simpler interpretation
+ Lower deviance
- Loses prediabetes information
- Binary classification only

Baseline Logit (Models 4 & 5):

+ Captures prediabetes stage
+ More comprehensive analysis
+ Three-category prediction
- Higher AIC (worse fit)
- More complex interpretation

Interesting Discovery

- BMI impact increases sharply with age (2.5% $\rightarrow$ 71% per unit).
- Gender effect reverses with age: young men lower risk, older men higher risk than women.
- **Poor physical health days increase diabetes risk only among those with very good/excellent general health.**
- Having both high cholesterol and heart disease doubles the odds of diabetes.
- Physical activity does not fully offset BMI: BMI effect slightly higher in active individuals.

Questions?

References I

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.
3. Winer, N., & Sowers, J. R. (2013). Epidemiology of diabetes. *Journal of Clinical Pharmacology*, 44(4), 397–405. https://doi.org/10.1177/0091270004263017