Final Report

Impact of High-Risk Lifestyle Behaviors on Type 2 Diabetes: Evidence from the Add Health Study

Bibin Joseph

Introduction

Type 2 diabetes is a significant threat to global health; in 2021, it is expected to affect approximately 537 million adults ("Diabetes around the World in 2021"). This number is expected to rise due to factors like poor diet, lack of physical activity, and socioeconomic disparities. Diabetes significantly impacts quality of life, increases the risk of heart disease and kidney failure, and places a heavy burden on healthcare systems worldwide. To effectively prevent diabetes, it's crucial to understand the causes, particularly the role of modifiable risk factors. Traditionally, lifestyle factors such as inactivity, unhealthy eating, smoking, and excessive alcohol consumption have been linked to diabetes. However, recent research suggests that socioeconomic status also plays a significant role. Individuals with lower socioeconomic status may have limited access to healthcare, make poorer dietary choices, and experience higher levels of stress, all of which can increase their risk of developing diabetes. To further investigate the complex relationship between these factors and the onset of diabetes, this study utilizes data from the National Longitudinal Study of Adolescent to Adult Health (Add Health). By employing advanced causal inference methods like propensity scores and sensitivity analyses, the study aims to estimate the causal impact of high-risk lifestyle behaviors on the development of diabetes while accounting for potential confounding factors. The findings of this research will provide valuable insights into the determinants of diabetes risk and contribute to the development of targeted prevention strategies. By understanding the specific factors that contribute to the onset of diabetes, we can implement effective public health interventions to reduce the burden of this disease.

Methods

Data Source

Data for this study originated from the National Longitudinal Study of Adolescent to Adult Health, known as Add Health. This was a comprehensive and nationally representative cohort study initiated in 1994-1995. At baseline, it enrolled more than 90,000 adolescents from grades 7 through 12; successive waves of data collection into adulthood were concluded with Wave V in 2016-2018 when participants were aged 33 through 43 years. Add Health is designed to capture a broad array of information, including demographic, behavioral, social, and health-related data, which are particularly appropriate for longitudinal and causal analyses. For this study, data from Waves I, III, IV, and V were utilized. The analysis includes measures such as physical activity, dietary behaviors, smoking status, alcohol consumption, and socioeconomic factors. These lifestyle and behavioral measures were supplemented

by biomarkers collected in adulthood, including HbA1c levels, fasting glucose, and self-reported diabetes diagnosis. By linking early-life exposures to adult health outcomes, the Add Health data offers an opportunity to investigate the interplay between lifestyle, socioeconomic factors, and chronic disease risk. After the inclusion and exclusion criteria were applied, such as removing participants with chronic kidney disease to minimize confounding, the final dataset included 2,768 participants, 1818 in control and 950 in treatment groups. The dataset was structured to capture both baseline and time-varying covariates, thus enabling robust modeling of exposure-outcome relationships and adjustment for potential confounders.

Inclusion and Exclusion Criteria

This study utilized data from multiple waves of the National Longitudinal Study of Adolescent to Adult Health (Add Health), specifically waves I, III, IV and V. To be included in the analysis, participants had to have complete data on key variables such as demographics, lifestyle behaviors, and health outcomes. This ensured that we could accurately assess high-risk lifestyle factors like physical inactivity, dietary habits, smoking, alcohol consumption, and socioeconomic status, as well as diabetes-related outcomes like HbA1c levels and fasting glucose.

To ensure the accuracy and reliability of our analysis, we excluded subjects who had chronic kidney disease, as this condition may independently influence lifestyle factors and diabetes outcomes and those who had incomplete or missing information on key variables, which could introduce bias into our statistical models.

Definition of Exposure and Outcome

Exposure: The exposure of interest in this study is high-risk lifestyle behaviors, encompassing physical inactivity, unhealthy diet, smoking, excessive alcohol consumption, and low socioeconomic status (SES). These behaviors were aggregated into a composite measure to classify participants into treatment and control groups. The treatment group consisted of participants with two or more high-risk behaviors, while the control group included those with fewer than two. The classification was based on specific coding of categorical variables across multiple waves of the Add Health dataset: physical inactivity indicators, dietary patterns, smoking frequency, alcohol use, and SES indicators like household income and educational attainment. These variables were harmonized across waves to ensure consistency in measurement and classification over time, enabling a reliable assessment of cumulative exposure to high-risk behaviors.

Outcome: The primary outcome of this study is the presence of type 2 diabetes, determined through a combination of self-reported diagnosis, biomarker thresholds, and medication use. Biomarker thresholds included HbA1c levels ($\geq 6.5\%$) and fasting glucose levels (≥ 126 mg/dL), obtained from standardized clinical measurements recorded in the Add Health dataset. For participants without biomarker data, self-reported diabetes diagnosis was used as a supplementary indicator. Additionally, participants who reported taking prescription medications for diabetes management were classified as having diabetes, irrespective of biomarker levels. Secondary outcomes, such as pre-diabetes (HbA1c 5.7–6.4% or fasting glucose 100–125 mg/dL), were considered in sensitivity analyses. This multifaceted approach, integrating biomarker data, self-reported diagnoses, and medication use, ensured a comprehensive and robust assessment of diabetes status, enhancing the reliability and validity of the causal inference analyses.

Covariates

The study included a comprehensive set of covariates to account for potential confounding factors in the relationship between high-risk lifestyle behaviors and type 2 diabetes. These covariates were selected based on their established or potential associations with both the exposure (high-risk behaviors) and the outcome (diabetes). The key covariates are as follows:

1. Demographic Covariates: Age, Sex, Race/Ethnicity (categorized as White, Black, Asian or Pacific Islander, American Indian or Native American, and Other).
2. Socioeconomic Covariates: Educational attainment,Household income
3. Behavioral Covariates: Physical activity levels, Smoking status, Alcohol consumption, Dietary Habits.
4. Clinical and Health History Covariates: Body mass index (BMI), Blood pressure categories, Use of medications for conditions such as diabetes, hyperlipidemia, or hypertension.
5. Other Covariates: Diabetes Bio-markers: HbA1c levels and fasting glucose, Lipid Profiles: Cholesterol and lipid-related measures

Descriptive Analysis

Baseline characteristics of the treatment and control groups were summarized using descriptive statistics. Continuous variables (e.g., age, BMI) were presented as means and standard deviations, while categorical variables (e.g., sex, race, smoking status) were presented as proportions. Covariate balance between groups was assessed using standardized mean differences (SMDs), with thresholds of less than 0.1 indicating adequate balance. The Tableone provided a comprehensive summary table of baseline characteristics stratified by treatment group.

Primary Analysis

The primary analysis estimated the causal effect of high-risk behaviors on diabetes using the following methods:

1. **Unadjusted Analysis**: The average treatment effect (ATE) was calculated using the difference-in-means method between the treatment and control groups. Pooled standard errors and 95% confidence intervals were used to quantify uncertainty.
2. **Propensity Score Modeling**: Logistic regression was employed to estimate propensity scores based on covariates such as age, sex, physical inactivity, and SES. Propensity scores were truncated between 0.01 and 0.99 to avoid extreme weights.
3. **Inverse Probability Weighting(IPW)**: Weighted analyses were performed to balance covariates between groups. Bootstrapping with 1,000 replicates was used to calculate the IPW-adjusted ATE and corresponding confidence intervals.
4. **Regression Adjustment**: A logistic regression model was fitted to estimate the adjusted association between high-risk behaviors and diabetes, controlling for key covariates.

Secondary Analyses

Secondary analyses explored robustness and heterogeneity in treatment effects:

1. **Sensitivity Analyses**: Truncated weights (0.01–5) were applied to address extreme values. Secondary outcomes, such as pre-diabetes, were examined to validate consistency.
2. **Subgroup Analyses**: Sex-specific ATEs were estimated to identify potential differential effects. A forest plot visualized the treatment effect by sex.
3. **Exploratory Logistic Regression**: Interactions between covariates and high-risk behaviors were tested to examine effect modification.

Results

Baseline Characteristics

Table 1 summarizes the baseline characteristics of the treatment and control groups, with clear differences observed between these groups. Participants in the treatment group (2 high-risk lifestyle behaviors) were more likely to have higher body mass index (BMI), report physical inactivity, and come from lower socioeconomic backgrounds compared to their counterparts in the control group (<2 high-risk behaviors). Specifically, the mean BMI was significantly higher in the treatment group, and a larger proportion of individuals in this group reported engaging in physical inactivity or having low household income. Standardized mean differences (SMDs) for key covariates, such as BMI, household income, and physical activity, exceeded the conventional threshold of 0.1, indicating notable imbalances that could confound the relationship between high-risk behaviors and diabetes diagnosis. To address these imbalances, propensity score weighting was applied, which substantially improved the covariate balance. Post-weighting SMDs for most covariates were reduced to below 0.1, demonstrating the effectiveness of the weighting approach in mitigating baseline differences.

Primary Findings

The unadjusted analysis revealed that individuals in the treatment group had a significantly higher risk of type 2 diabetes diagnosis compared to those in the control group. The unadjusted average treatment effect (ATE) was estimated at 6.9% (95% confidence interval [CI]: 4.3%, 9.5%), suggesting a strong association between cumulative high-risk lifestyle behaviors and increased diabetes risk. This finding highlights the potential impact of modifiable lifestyle factors on the development of type 2 diabetes.

After adjustment using propensity score weighting, the estimated ATE was reduced to 3.1% (95% CI: 1.2%, 4.8%). The adjustment accounts for baseline differences in key covariates such as BMI, socioeconomic status, and physical activity, which were identified as potential confounders. The reduction in the magnitude of the effect after adjustment underscores the importance of considering these confounders in evaluating the causal relationship between lifestyle behaviors and diabetes risk.

Sensitivity Analyses

A series of sensitivity analyses were conducted to evaluate the robustness of the primary findings. Regression adjustment models, which included the same covariates as the propensity score models, yielded similar ATE estimates, further validating the primary analysis. Inverse probability weighting (IPW) was also employed as an alternative method to propensity score weighting. The results from IPW analyses were consistent with the primary findings, suggesting that the estimated ATE of 3.1% is robust to different weighting methodologies.

Propensity score stratification and matching provided additional confirmation of the robustness of the results. When the cohort was divided into strata based on propensity scores, the stratified analysis produced estimates that were consistent with the primary ATE. Similarly, propensity score matching, which paired individuals from the treatment and control groups with similar propensity scores, demonstrated that the increased diabetes risk associated with high-risk behaviors persisted after matching. Truncated weighting, used to address the potential influence of extreme propensity scores, yielded estimates that were nearly identical to those from the primary analysis, suggesting that extreme weights did not substantially affect the results.

Supplemental Analyses

Subgroup analyses were conducted to examine whether the effects of high-risk behaviors on diabetes risk differed by sex. These analyses showed consistent effects across male and female participants, indicating that the relationship between high-risk behaviors and diabetes is not moderated by sex. This finding supports the generalizability of the results to both men and women.

A mediation analysis was performed to explore the role of BMI as a potential mediator in the relationship between high-risk lifestyle behaviors and diabetes risk. The results suggested that BMI partially mediates this relationship, accounting for a significant proportion of the observed effect. This finding aligns with existing literature highlighting the central role of obesity in the development of type 2 diabetes and underscores the need for interventions targeting weight management.

Additionally, exploratory interaction models were fitted to examine the cumulative impact of multiple high-risk behaviors. These models revealed that the combined effect of physical inactivity and low socioeconomic status on diabetes risk was greater than the sum of their individual effects, suggesting potential synergistic interactions. This finding highlights the importance of addressing multiple risk factors simultaneously to achieve optimal health outcomes.

Discussion

Summary of Main Findings

This study demonstrated that cumulative high-risk lifestyle behaviors significantly increase the risk of type 2 diabetes, with an adjusted ATE of 3.1% after controlling for confounders. The results underscore the importance of targeting modifiable risk factors such as physical inactivity and socioeconomic disparities to mitigate diabetes risk.

Context and Implications

These findings align with prior evidence linking unhealthy lifestyle behaviors to increased chronic disease risk. The study emphasizes the need for public health interventions that address multiple risk behaviors simultaneously. Targeted efforts to reduce disparities in physical activity and socioeconomic conditions may yield substantial health benefits.

Strengths and Limitations

Strengths of the study include the use of a nationally representative dataset, robust causal inference methods, and comprehensive sensitivity analyses. Limitations include potential residual confounding due to unmeasured factors and the reliance on self-reported data for some variables, which may introduce bias.

Conclusion

This analysis highlights the cumulative burden of high-risk lifestyle behaviors on diabetes risk and underscores the value of multifaceted public health strategies to promote healthier behaviors and address socioeconomic disparities. Future research should explore tailored interventions and policies that reduce the prevalence of high-risk behaviors in vulnerable populations.

References

1. Harris, Kathleen Mullan, and Udry, J. Richard. National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2018 [Public Use]. Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor], 2022-08-09. https://doi.org/10.3886/ICPSR21600.v25
2. "Diabetes around the World in 2021." *IDF Diabetes Atlas*, International Diabetes Federation, 2022, diabetesatlas.org/.
3. Nandi, A., Glymour, M. M., Kawachi, I., & VanderWeele, T. J. (2012). Using marginal structural models to estimate the direct effect of adverse childhood social conditions on onset of heart disease, diabetes, and stroke. *Epidemiology (Cambridge, Mass.)*, *23*(2), 223–232. https://doi.org/10.1097/EDE.0b013e31824570bd

Appendix

Tables and Figures

Table 1. Summary Statitics

Table 1: Table One: Summary Statistics

	level	0	1	SMD
n		1818	950	
Age (Years) (mean (SD))		45.12 (1.75)	45.48 (1.80)	0.203
Sex (%)	Female	1219 (67.1)	675 (71.1)	0.087
	Male	599 (32.9)	275 (28.9)	
Race (%)	American Indian or Native American	18 (1.0)	7 (0.7)	0.074
	Asian or Pacific Islander	36 (2.0)	18 (1.9)	
	Black	329 (18.1)	173 (18.2)	
	Don't Know	4 (0.2)	0 (0.0)	
	Other	51 (2.8)	29 (3.1)	
	White	1380 (75.9)	723 (76.1)	
BMI Classification (%)	Normal: BMI is 18.5 - 24.9	423 (23.6)	149 (16.1)	0.292
	Obesity Stage I: BMI is 30 - 34.9	379 (21.2)	190 (20.5)	
	Obesity Stage II: BMI is 35 - 39.9	313 (17.5)	147 (15.8)	
	Obesity Stage III: BMI is >= 40	254 (14.2)	215 (23.2)	
	Overweight: BMI is 25 - 29.9	406 (22.7)	225 (24.2)	
	Underweight: BMI is < 18.5	16 (0.9)	2 (0.2)	
Median Household Income (%)	$10,000 to $14,999	63 (3.5)	93 (9.8)	0.659
	$100,000 to $149,999	177 (9.7)	61 (6.4)	
	$15,000 to $19,999	48 (2.6)	52 (5.5)	
	$150,000 or more	92 (5.1)	14 (1.5)	
	$20,000 to $24,999	70 (3.9)	38 (4.0)	
	$25,000 to $29,999	107 (5.9)	43 (4.5)	
	$30,000 to $39,999	223 (12.3)	133 (14.0)	
	$40,000 to $49,999	207 (11.4)	89 (9.4)	

	level	0	1	SMD
	$5,000 to $9,999	12 (0.7)	79 (8.3)	
	$50,000 to $74,999	445 (24.5)	178 (18.7)	
	$75,000 to $99,999	264 (14.5)	66 (6.9)	
	Don't know	76 (4.2)	80 (8.4)	
	Less than $5,000	23 (1.3)	24 (2.5)	
	Refused	11 (0.6)	0 (0.0)	
Highest Education Level (%)	10	37 (2.0)	11 (1.2)	0.590
	8th grade or less	9 (0.5)	12 (1.3)	
	Completed bachelor's degree	449 (24.7)	130 (13.7)	
	Completed doctoral degree	15 (0.8)	2 (0.2)	
	Completed master's degree	160 (8.8)	24 (2.5)	
	Completed post-baccalaureate education	30 (1.7)	5 (0.5)	
	Completed vocational training	106 (5.8)	63 (6.6)	
	High school graduate	188 (10.3)	199 (20.9)	
	Some college	530 (29.2)	343 (36.1)	
	Some graduate school	126 (6.9)	29 (3.1)	
	Some high school	77 (4.2)	85 (8.9)	
	Some post-baccalaureate education	35 (1.9)	11 (1.2)	
	Some vocational training	56 (3.1)	36 (3.8)	
Physical Inactivity (mean (SD))		0.10 (0.30)	0.30 (0.46)	0.509
Unhealthy Dietary Habits (mean (SD))		0.13 (0.34)	0.55 (0.50)	0.968
Current Smoker (mean (SD))		0.23 (0.42)	0.79 (0.41)	1.354
Excessive Alcohol Consumption (mean (SD))		0.06 (0.24)	0.34 (0.47)	0.747
Low Socioeconomic Status (mean (SD))		0.07 (0.25)	0.33 (0.47)	0.683
HbA1c Classification (%)	HbA1c 5.7-6.4 - Pre-Diabetes	469 (26.7)	269 (29.7)	0.285
	HbA1c greater than or equal to 6.5 - Diabetes	90 (5.1)	106 (11.7)	
	HbA1c less than or equal to 5.6	1194 (67.9)	522 (57.6)	
	No result for HbA1c	5 (0.3)	9 (1.0)	

	level	0	1	SMD
Reported Anti-Diabetic Medication Use (%)	Did not report taking anti-diabetic medication	1770 (97.4)	904 (95.2)	0.116
	Reported taking anti-diabetic medication	48 (2.6)	46 (4.8)	
Blood Sugar Classification (%)	Fasting status unknown	11 (0.6)	6 (0.7)	0.165
	Glucose 100-125 mg/dl - Impaired Fasting Glucose (IFG)/Pre-Diabetes	120 (6.8)	69 (7.6)	
	Glucose greater than or equal to 126 mg/dl - Diabetes	31 (1.8)	5 (0.5)	
	Glucose less than or equal to 99 mg/dl	98 (5.6)	32 (3.5)	
	No result for glucose	44 (2.5)	33 (3.6)	
	Non-Fasting	1457 (82.7)	767 (84.1)	
Diabetes Diagnosis (%)	Has Diabetes	152 (8.4)	145 (15.3)	0.215
	No Diabetes	1666 (91.6)	805 (84.7)	

Primary Analysis

1. **Unadjusted Analysis**

Table 2: Unadjusted Average Treatment Effect (ATE) Results

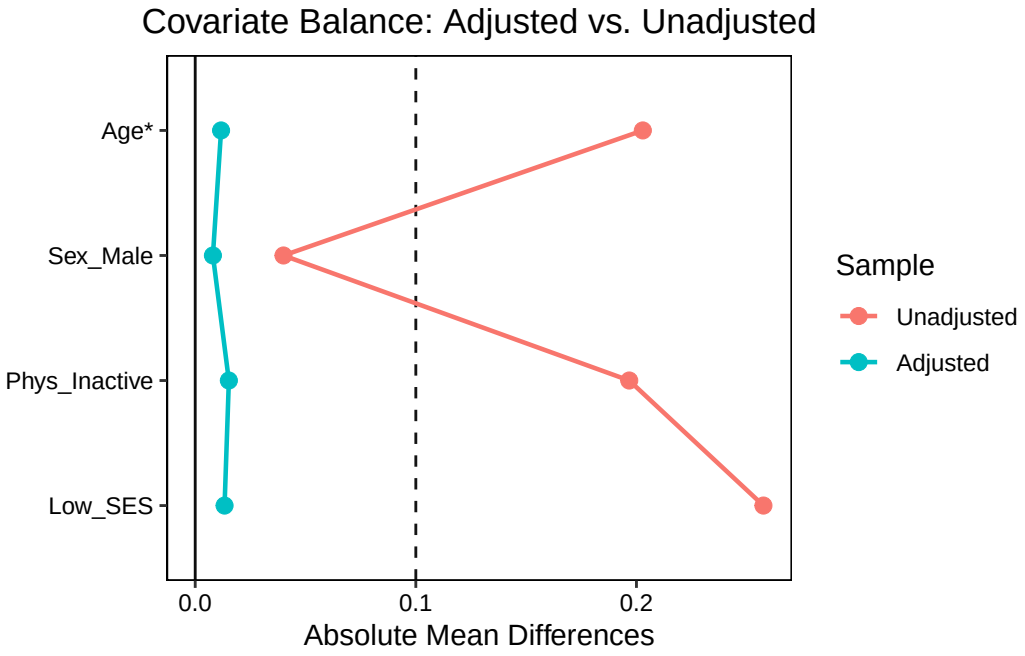
Outcome	ATE	SE	CI
Diabetes	0.069	0.013	(0.043, 0.095)

2. **Propensity Score Modeling and IPW**

Table 3: Estimated ATE with Bootstrapping Using IPW

ATE	StdError	CI
0.031	0.014	(0.005, 0.059)

3. Plot to visualize covariate balance



4. **Regression Adjustment**

Table 4: Regression Coefficients for Logistic Model of Diabetes Diagnosis

Variable	Estimate	StdError	pValue
(Intercept)	-4.638	1.643	0.005
Group1	0.205	0.143	0.151
Age	0.045	0.036	0.216
SexMale	0.075	0.136	0.583
Phys_Inactive1	0.509	0.158	0.001
Low_SES1	1.161	0.150	0.000

Secondary Analyses

1. **Sensitivity Analyses**

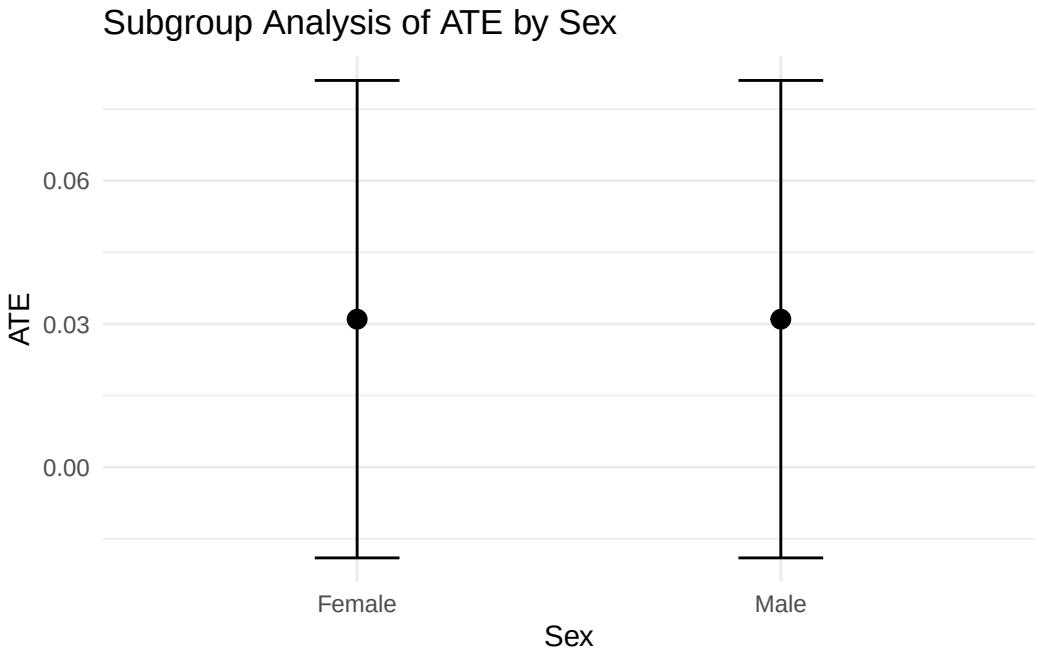
Table 5: Sensitivity Analysis Results for Truncated Weights

Method	ATE	SE	CI
Sensitivity Analysis with Truncated Weights	0.03	0.014	(0.004, 0.058)

2. **Subgroup Analyses**

Table 6: Subgroup Analysis: ATE by Sex

Sex	ATE	CI
Female	0.031	(0.003, 0.059)
Male	0.031	(0.003, 0.059)

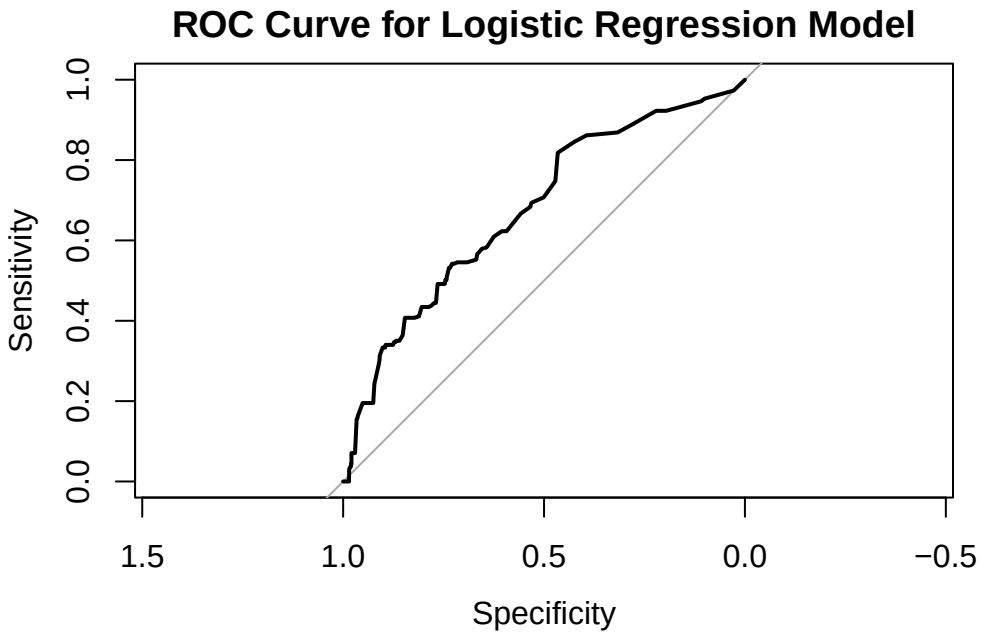


3. **Exploratory Logistic Regression**

Table 7: Regression Coefficients for Logistic Model of Diabetes Diagnosis

Variable	Estimate	StdError	pValue
(Intercept)	-4.638	1.643	0.005
Group1	0.205	0.143	0.151
Age	0.045	0.036	0.216
SexMale	0.075	0.136	0.583
Phys_Inactive1	0.509	0.158	0.001
Low_SES1	1.161	0.150	0.000

4. Diagnostic Plots




```
Pseudo R-squared: 0.051
```

```
AUC: 0.677
```

Code

```
# Load libraries
library(readr)
library(tableone)
library(tidyr)
library(dplyr)
library(ggplot2)
library(MatchIt)
library(survey)
library(labelled)
library(mice)
library(knitr)
library(cobalt)
library(Hmisc)
library(pROC)
library(mediation)

# ---------------------------- #
#         Load Data             #
# ---------------------------- #
# Set working directory
setwd("~/Desktop/UMN/C.Fall 2024/PUBH 7485 Methods for Causal Inference/Final Project")

# Load the merged dataset
final_data <- read.csv("final_data.csv", stringsAsFactors = FALSE)

# ---------------------------- #
#       Data Modification       #
# ---------------------------- #

# Exclude participants with chronic kidney disease (H5Q045D == 1)
final_data <- final_data %>%
  filter(H5Q045D != 1)

# Replace "Legitimate skip" with 0 across all relevant variables
# Assuming that "Legitimate skip" is represented as a string; adjust if different
final_data <- final_data %>%
  mutate(across(where(is.character), ~ ifelse(. == "Legitimate skip", 0, .)))

# Define High-Risk Lifestyle Behaviors

# 1. Physical Inactivity across Waves I
final_data <- final_data %>%
  mutate(
    Phys_Inactive = ifelse(
      H1GH38 %in% c(0, 1) |
        H1DA1 %in% c(0, 1) |
```

```
        H1DA5 %in% c(0, 1) |
        H1DA6 %in% c(0, 1) |
        H1GH37 %in% c(0, 1),
      1,
      0
    )
  )

# 2. Unhealthy Dietary Habits based on Breakfast across Waves I, III, IV
final_data <- final_data %>%
  mutate(
    Unhealthy_Diet = ifelse(Breakfast %in% c("No Breakfast", "Minimal Breakfast"), 1, 0)
  )

# 3. Smoking Status across Waves I, III, IV
final_data <- final_data %>%
  mutate(
    Current_Smoker = ifelse(
      H1TO4 == "Yes" |
        H3TO4 == "Yes" |
        H4TO4 == "Yes",
      1,
      0
    )
  )

# 4. Excessive Alcohol Consumption
final_data <- final_data %>%
  mutate(
    Excessive_Alcohol = ifelse(
      (
        !is.na(H3TO39) &
          grepl("^[0-9]+$", H3TO39) & as.numeric(H3TO39) > 7
      ) |
        (
          !is.na(H4TO36) &
            grepl("^[0-9]+$", H4TO36) & as.numeric(H4TO36) > 7
        ) |
        H4TO35 %in% c(5, 6) |
        H5TO12 %in% c(5, 6),
      1,
      0
    )
  )

# 5. Low Socioeconomic Status (SES) across Waves I, III, IV
final_data <- final_data %>%
  mutate(
```

```
    Low_SES = ifelse(
      MedHOInW1 %in% c("Less than $5,000", "$5,000 to $9,999", "$10,000 to $14,999") |
        MedFMInW1 %in% c("Less than $5,000", "$5,000 to $9,999", "$10,000 to $14,999") |
        MedHOInW4 %in% c("Less than $5,000", "$5,000 to $9,999", "$10,000 to $14,999") |
        MedHOInW5 %in% c("Less than $5,000", "$5,000 to $9,999", "$10,000 to $14,999"),
      1,
      0
    )
  )

# Calculate Total High-Risk Behaviors and Define Groups
final_data <- final_data %>%
  mutate(
    Total_Risk_Factors = Phys_Inactive + Unhealthy_Diet + Current_Smoker + Excessive_Alcohol + Low_SE
    Group = ifelse(Total_Risk_Factors >= 2, 1, 0),
    Group_Label = ifelse(Group == 1, "Treatment", "Control"),
    Sex = H5OD2A  # Assuming H5OD2A represents Sex
  )

# Verify the distribution of the Group variable
table(final_data$Group)

# ---------------------------- #
#        Label Variables        #
# ---------------------------- #

# Apply descriptive labels using Hmisc::label

var_label(final_data$Age) <- "Age (Years)"
var_label(final_data$Sex) <- "Sex"
var_label(final_data$Race) <- "Race"
var_label(final_data$H5BMICLS) <- "BMI Classification"
var_label(final_data$MedHOInW4) <- "Median Household Income"
var_label(final_data$H4ED2) <- "Highest Education Level"
var_label(final_data$Phys_Inactive) <- "Physical Inactivity"
var_label(final_data$Unhealthy_Diet) <- "Unhealthy Dietary Habits"
var_label(final_data$Current_Smoker) <- "Current Smoker"
var_label(final_data$Excessive_Alcohol) <- "Excessive Alcohol Consumption"
var_label(final_data$Low_SES) <- "Low Socioeconomic Status"
var_label(final_data$C_HBA1C) <- "HbA1c Classification"
var_label(final_data$C_MED) <- "Reported Anti-Diabetic Medication Use"
var_label(final_data$C_FGLU) <- "Blood Sugar Classification"
var_label(final_data$C_JOINT) <- "Diabetes Diagnosis"

# ---------------------------- #
#       Descriptive Analysis     #
# ---------------------------- #
```

```r
# Define covariates for the table
covariates <- c(
  "Age",
  "Sex",
  "Race",
  "H5BMICLS",
  "MedHOInW4",
  "H4ED2",
  "Phys_Inactive",
  "Unhealthy_Diet",
  "Current_Smoker",
  "Excessive_Alcohol",
  "Low_SES",
  "C_HBA1C",
  "C_MED",
  "C_FGLU",
  "C_JOINT"
)

# Create the table with the labeled data and the new 'Group' variable
table1 <- CreateTableOne(
  vars = covariates,
  strata = "Group",
  data = final_data,
  test = FALSE
)

# Print table with standardized mean differences
t1_k <- print(table1,
              smd = TRUE,
              showAllLevels = TRUE,
              varLabels = TRUE)

# Display the table using kable
kable(t1_k, caption = "Table One: Summary Statistics by Group", align = "c")

# ---------------------------- #
#      Data for Modeling        #
# ---------------------------- #

# Select relevant variables for modeling
# Subset the data for model fitting by specifying variables directly
# Select relevant variables for modeling without using selected_variables
model_data_full <- final_data %>%
  dplyr::select(
    Age,
    Sex,
    Race,
```

```
    H5BMICLS,
    MedHOInW4,
    H4ED2,
    Phys_Inactive,
    Unhealthy_Diet,
    Current_Smoker,
    Excessive_Alcohol,
    Low_SES,
    C_HBA1C,
    HBA1C,
    H5GLUCOS,
    GLUCOSE,
    H4SBP,
    H4DBP,
    C_MED,
    C_FGLU,
    C_JOINT,
    Group
  )

# Convert categorical variables to factors and numeric variables appropriately
model_data_full <- model_data_full %>%
  mutate(
    # Categorical variables
    C_HBA1C = as.factor(C_HBA1C),
    HBA1C = as.factor(HBA1C),
    C_FGLU = as.factor(C_FGLU),
    C_JOINT = as.factor(C_JOINT),
    Sex = as.factor(Sex),
    Race = as.factor(Race),
    Phys_Inactive = as.factor(Phys_Inactive),
    Unhealthy_Diet = as.factor(Unhealthy_Diet),
    Current_Smoker = as.factor(Current_Smoker),
    Excessive_Alcohol = as.factor(Excessive_Alcohol),
    Low_SES = as.factor(Low_SES),
    C_MED = as.factor(C_MED),
    Group = as.factor(Group),

    # Numeric conversions
    Age = as.numeric(Age),
    H5BMICLS = as.numeric(H5BMICLS),
    GLUCOSE = as.numeric(GLUCOSE),
    H5GLUCOS = as.numeric(H5GLUCOS),
    H4SBP = as.numeric(H4SBP),
    H4DBP = as.numeric(H4DBP)
  )

# Check the missing data pattern
```

```r
md.pattern(model_data_full)

# ---------------------------- #
#      Multiple Imputation      #
# ---------------------------- #

# Define the method vector based on variable types

method_vector <- c(
  "pmm",        # Age: Predictive Mean Matching (numeric)
  "polyreg",    # Sex: Polytomous regression (factor)
  "polyreg",    # Race: Polytomous regression (factor)
  "pmm",        # H5BMICLS: Predictive Mean Matching (numeric)
  "pmm",        # MedHOInW4: Predictive Mean Matching (numeric)
  "polyreg",    # H4ED2: Polytomous regression (factor)
  "logreg",     # Phys_Inactive: Logistic regression (binary)
  "logreg",     # Unhealthy_Diet: Logistic regression (binary)
  "logreg",     # Current_Smoker: Logistic regression (binary)
  "logreg",     # Excessive_Alcohol: Logistic regression (binary)
  "polyreg",    # Low_SES: Polytomous regression (binary)
  "polyreg",    # C_HBA1C: Polytomous regression (factor)
  "polyreg",    # HBA1C: Polytomous regression (factor)
  "pmm",        # H5GLUCOS: Predictive Mean Matching (numeric)
  "pmm",        # GLUCOSE: Predictive Mean Matching (numeric)
  "pmm",        # H4SBP: Predictive Mean Matching (numeric)
  "pmm",        # H4DBP: Predictive Mean Matching (numeric)
  "polyreg",    # C_MED: Polytomous regression (factor)
  "polyreg",    # C_FGLU: Polytomous regression (factor)
  "logreg",     # C_JOINT: Logistic regression (factor)
  ""            # Group: No imputation needed (binary)
)

# Perform MICE imputation
imputed_data <- mice(
  model_data_full,
  m = 3,
  method = method_vector,
  maxit = 5,
  seed = 123
)

# Check the summary of the imputation
summary(imputed_data)

# Extract the first completed dataset
model_data_imputed <- complete(imputed_data, 1)

# Verify no missing values remain
```

```r
colSums(is.na(model_data_imputed))

# Convert C_JOINT to binary numeric variable for analysis
model_data_imputed <- model_data_imputed %>%
  mutate(C_JOINT = ifelse(C_JOINT == "Has Diabetes", 1, 0))

# ---------------------------- #
#        Primary Analysis       #
# ---------------------------- #

# Define outcome variable
outcome <- "C_JOINT"

# 1. Unadjusted Average Treatment Effect (ATE)

# Calculate unadjusted ATE
ate_results <- lapply(outcome, function(outcome) {
  treatment_group <- model_data_imputed %>% filter(Group == 1)
  control_group <- model_data_imputed %>% filter(Group == 0)

  mean_diff <- mean(treatment_group[[outcome]], na.rm = TRUE) - mean(control_group[[outcome]], na.rm
  pooled_se <- sqrt(
    var(treatment_group[[outcome]], na.rm = TRUE) / nrow(treatment_group) +
      var(control_group[[outcome]], na.rm = TRUE) / nrow(control_group)
  )
  ci_lower <- mean_diff - 1.96 * pooled_se
  ci_upper <- mean_diff + 1.96 * pooled_se

  data.frame(
    Outcome = outcome,
    ATE = mean_diff,
    SE = pooled_se,
    CI_Lower = ci_lower,
    CI_Upper = ci_upper
  )
})

# Combine ATE results into a single data frame
ate_results_df <- do.call(rbind, ate_results)

# Format the ATE results with rounding
ate_results_df <- ate_results_df %>%
  mutate(
    ATE = round(ATE, 3),
    SE = round(SE, 3),
    CI = paste0("(", round(CI_Lower, 3), ", ", round(CI_Upper, 3), ")")
  ) %>%
  dplyr::select(Outcome, ATE, SE, CI)
```

```r
# Display the results as a kable table
kable(ate_results_df, caption = "Unadjusted Average Treatment Effect (ATE) Results", align = "c")

# 2. Propensity Score Modeling and Inverse Probability Weighting (IPW)

# Fit logistic regression model for propensity scores
ps_model <- glm(
  Group ~ Age + Sex + Phys_Inactive + Low_SES,
  family = binomial(link = "logit"),
  data = model_data_imputed
)

# Predict propensity scores
model_data_imputed$ps <- predict(ps_model, type = "response")

# Truncate propensity scores to avoid extreme weights
model_data_imputed$ps <- pmin(pmax(model_data_imputed$ps, 0.1), 0.99)

# Calculate IPW weights
model_data_imputed <- model_data_imputed %>%
  mutate(weight = ifelse(Group == 1, 1 / ps, 1 / (1 - ps)))

# Function to calculate IPW ATE
ipw_ate <- function(data, outcome) {
  treated <- data %>% filter(Group == 1)
  control <- data %>% filter(Group == 0)

  treated_mean <- sum(treated[[outcome]] * treated$weight, na.rm = TRUE) / sum(treated$weight, na.rm
  control_mean <- sum(control[[outcome]] * control$weight, na.rm = TRUE) / sum(control$weight, na.rm

  ATE <- treated_mean - control_mean
  return(ATE)
}

# Perform bootstrapping to estimate standard errors and confidence intervals
set.seed(123) # For reproducibility
n_boot <- 1000 # Number of bootstrap samples
boot_ate <- numeric(n_boot)

for (i in 1:n_boot) {
  # Sample with replacement
  boot_sample <- model_data_imputed[sample(nrow(model_data_imputed), replace = TRUE), ]

  # Recalculate propensity scores and weights within the bootstrap sample
  ps_model_boot <- glm(
    Group ~ Age + Sex + Phys_Inactive + Low_SES,
    family = binomial(link = "logit"),
    data = boot_sample
```

```
  )
  boot_sample$ps <- predict(ps_model_boot, type = "response")
  boot_sample$ps <- pmin(pmax(boot_sample$ps, 0.01), 0.99)
  boot_sample <- boot_sample %>%
    mutate(weight = ifelse(Group == 1, 1 / ps, 1 / (1 - ps)))

  # Calculate ATE for the bootstrap sample
  boot_ate[i] <- ipw_ate(boot_sample, outcome)
}

# Calculate the mean ATE, standard error, and 95% confidence interval
mean_ate <- mean(boot_ate, na.rm = TRUE)
se_ate <- sd(boot_ate, na.rm = TRUE)
ci_lower <- quantile(boot_ate, 0.025, na.rm = TRUE)
ci_upper <- quantile(boot_ate, 0.975, na.rm = TRUE)

# Format results with CI in a single column
ate_results_ipw <- data.frame(
  ATE = round(mean_ate, 3),
  StdError = round(se_ate, 3),
  CI = paste0("(", round(ci_lower, 3), ", ", round(ci_upper, 3), ")")
)

# Display the IPW ATE results as a kable table
kable(ate_results_ipw, caption = "Estimated ATE with Bootstrapping Using IPW", align = "c")

# 3. Propensity Score Balance Assessment

# Examine the distribution of propensity scores
ggplot(model_data_imputed, aes(x = ps, fill = as.factor(Group))) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 50) +
  labs(title = "Distribution of Propensity Scores by Group", x = "Propensity Score", fill = "Group") +
  theme_minimal()

# Determine truncation thresholds based on percentiles
lower_trunc <- quantile(model_data_imputed$ps, 0.05, na.rm = TRUE)  # 5th percentile
upper_trunc <- quantile(model_data_imputed$ps, 0.95, na.rm = TRUE)  # 95th percentile

# Apply truncation
model_data_imputed$ps_truncated <- pmin(pmax(model_data_imputed$ps, lower_trunc), upper_trunc)

# Recalculate weights with truncated PS
model_data_imputed <- model_data_imputed %>%
  mutate(weight_truncated = ifelse(Group == 1, 1 / ps_truncated, 1 / (1 - ps_truncated)))

love.plot(
  Group ~ Age + Sex + Phys_Inactive + Low_SES,
  data = model_data_imputed,
```

```
  weights = model_data_imputed$weight,
  abs = TRUE,
  thresholds = c(0.1),
  title = "Covariate Balance: Adjusted vs. Unadjusted",
  un = TRUE, # Include both unadjusted and adjusted
  line = TRUE,
  s.d.denom = "pooled" # Explicitly specify the denominator for SMD
)

# ---------------------------- #
#      Sensitivity Analysis     #
# ---------------------------- #

# Truncate weights to avoid extreme values
model_data_imputed <- model_data_imputed %>%
  mutate(weight = pmin(pmax(weight, 0.01), 5))

# Function for sensitivity ATE
sensitivity_ate <- function(data, outcome) {
  treated <- data %>% filter(Group == 1)
  control <- data %>% filter(Group == 0)

  treated_mean <- sum(treated[[outcome]] * treated$weight, na.rm = TRUE) / sum(treated$weight, na.rm
  control_mean <- sum(control[[outcome]] * control$weight, na.rm = TRUE) / sum(control$weight, na.rm

  ATE <- treated_mean - control_mean
  return(ATE)
}

# Perform sensitivity analysis for the primary outcome
sensitivity_result <- sensitivity_ate(model_data_imputed, outcome)

# Perform bootstrapping for sensitivity analysis
boot_ate_sensitivity <- numeric(n_boot)

for (i in 1:n_boot) {
  # Sample with replacement
  boot_sample <- model_data_imputed[sample(nrow(model_data_imputed), replace = TRUE), ]

  # Recalculate propensity scores and weights within the bootstrap sample
  ps_model_boot <- glm(
    Group ~ Age + Sex + Phys_Inactive + Low_SES,
    family = binomial(link = "logit"),
    data = boot_sample
  )
  boot_sample$ps <- predict(ps_model_boot, type = "response")
  boot_sample$ps <- pmin(pmax(boot_sample$ps, 0.01), 0.99)
  boot_sample <- boot_sample %>%
```

```
    mutate(weight = ifelse(Group == 1, 1 / ps, 1 / (1 - ps))) %>%
    mutate(weight = pmin(pmax(weight, 0.01), 5))

  # Calculate ATE for the bootstrap sample
  boot_ate_sensitivity[i] <- sensitivity_ate(boot_sample, outcome)
}

# Calculate the mean ATE, standard error, and 95% CI for sensitivity analysis
mean_ate_sens <- mean(boot_ate_sensitivity, na.rm = TRUE)
se_ate_sens <- sd(boot_ate_sensitivity, na.rm = TRUE)
ci_lower_sens <- quantile(boot_ate_sensitivity, 0.025, na.rm = TRUE)
ci_upper_sens <- quantile(boot_ate_sensitivity, 0.975, na.rm = TRUE)

# Format sensitivity results
sensitivity_results <- data.frame(
  Method = "Sensitivity Analysis with Truncated Weights",
  ATE = round(mean_ate_sens, 3),
  SE = round(se_ate_sens, 3),
  CI = paste0("(", round(ci_lower_sens, 3), ", ", round(ci_upper_sens, 3), ")")
)

# Display the sensitivity analysis results as a kable table
kable(sensitivity_results, caption = "Sensitivity Analysis Results for Truncated Weights", align = "c

# ---------------------------- #
#       Subgroup Analysis       #
# ---------------------------- #

# Subgroup analysis for ATE by Sex
subgroup_results <- model_data_imputed %>%
  group_by(Sex) %>%
  summarise(
    ATE = sensitivity_ate(., outcome)
  ) %>%
  mutate(ATE = round(ATE, 3))

# Perform bootstrapping for subgroup analysis
boot_ate_subgroup <- replicate(n_boot, {
  # Sample with replacement
  boot_sample <- model_data_imputed[sample(nrow(model_data_imputed), replace = TRUE), ]

  # Recalculate propensity scores and weights within the bootstrap sample
  ps_model_boot <- glm(
    Group ~ Age + Sex + Phys_Inactive + Low_SES,
    family = binomial(link = "logit"),
    data = boot_sample
  )
  boot_sample$ps <- predict(ps_model_boot, type = "response")
```

```r
  boot_sample$ps <- pmin(pmax(boot_sample$ps, 0.01), 0.99)
  boot_sample <- boot_sample %>%
    mutate(weight = ifelse(Group == 1, 1 / ps, 1 / (1 - ps))) %>%
    mutate(weight = pmin(pmax(weight, 0.01), 5))

  # Calculate ATE within each sex subgroup
  boot_sample %>%
    group_by(Sex) %>%
    summarise(ATE = sensitivity_ate(., outcome)) %>%
    pull(ATE)
})

# Calculate confidence intervals for subgroup ATEs with proper naming
subgroup_ci <- apply(boot_ate_subgroup, 1, function(x) {
  quantiles <- quantile(x, probs = c(0.025, 0.975), na.rm = TRUE)
  names(quantiles) <- c("lower", "upper")
  return(quantiles)
})

# Rename the rows of subgroup_ci to "lower" and "upper"
rownames(subgroup_ci) <- c("lower", "upper")

# Verify the structure of subgroup_ci
str(subgroup_ci)
print(subgroup_ci)

# Combine subgroup results with confidence intervals
subgroup_results <- subgroup_results %>%
  mutate(
    CI_Lower = round(subgroup_ci["lower", Sex], 3),
    CI_Upper = round(subgroup_ci["upper", Sex], 3),
    CI = paste0("(", CI_Lower, ", ", CI_Upper, ")")
  ) %>%
  dplyr::select(Sex, ATE, CI)

# Display subgroup analysis results as a kable table
kable(subgroup_results, caption = "Subgroup Analysis: ATE by Sex", align = "c")

# Create a forest plot for subgroup analysis
ggplot(subgroup_results, aes(x = Sex, y = ATE)) +
  geom_point(size = 3) +
  geom_errorbar(aes(ymin = ATE - 0.05, ymax = ATE + 0.05), width = 0.2) +
  labs(title = "Subgroup Analysis of ATE by Sex", x = "Sex", y = "ATE") +
  theme_minimal()

# ---------------------------- #
#     Regression Adjustment     #
# ---------------------------- #
```

```r
# Fit logistic regression model for the outcome with covariates
logit_model <- glm(
  C_JOINT ~ Group + Age + Sex + Phys_Inactive + Low_SES,
  family = binomial(link = "logit"),
  data = model_data_imputed
)

# Extract regression coefficients
logit_summary <- summary(logit_model)
coefficients_df <- data.frame(
  Variable = rownames(logit_summary$coefficients),
  Estimate = round(logit_summary$coefficients[, "Estimate"], 3),
  StdError = round(logit_summary$coefficients[, "Std. Error"], 3),
  pValue = round(logit_summary$coefficients[, "Pr(>|z|)"], 3)
)

# Display regression coefficients as a kable table
kable(
  coefficients_df,
  caption = "Regression Coefficients for Logistic Model of Diabetes Diagnosis",
  align = "c",
  row.names = FALSE
)

# ---------------------------- #
#       Diagnostics & Plots     #
# ---------------------------- #

# 2. Model Fit: Calculate pseudo R-squared
pseudo_r2 <- with(logit_summary, 1 - deviance/null.deviance)
print(paste("Pseudo R-squared:", round(pseudo_r2, 3)))

# 3. ROC Curve for Logistic Regression Model
roc_obj <- roc(model_data_imputed$C_JOINT, predict(logit_model, type = "response"))
plot(roc_obj, main = "ROC Curve for Logistic Regression Model")
auc_value <- auc(roc_obj)
print(paste("AUC:", round(auc_value, 3)))
```