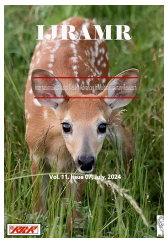
ISSN : 2350-0743

International Journal of Recent Advances in Multidisciplinary Research

Vol. 11, Issue 07, pp. 10135-10150, July, 2024



RESEARCH ARTICLE

DECODING DIABETES: UNDERSTANDING DIAGNOSIS THROUGH MEDICATION PRESCRIPTIONS

***Turgud Valiyev and Ulviyya Abasova**

Faculty of Economics, University of Warsaw, Warsaw, Poland

ARTICLE INFO

Article History

Received 10th April, 2024
Received in revised form
16th May, 2024
Accepted 17th June, 2024
Published online 30th July, 2024

Keywords:

Diabetes Diagnosis, Logistic Regression, Healthcare Utilization, Demographic Factors, Clinical Interventions.

****Corresponding author: Turgud Valiyev***

ABSTRACT

This study's analysis of a large dataset using logistic regression reveals important factors influencing diabetes diagnosis that are influenced by clinical history, healthcare consumption, and demography. The results show that older persons had a higher risk of developing diabetes, with notable variations observed in age groups, ethnicities, and genders. Notably, normal glucose levels lower the chance of diabetes, whereas older age, non-Caucasian ethnicity, and greater emergency healthcare utilization correspond with a higher risk. Given the complex interactions between numerous variables, these findings highlight the necessity of focused healthcare policies and initiatives for the optimal management and prevention of diabetes. This study provides insightful advice for improving clinical procedures and creating reasonable health policy.

Citation: Turgud Valiyev and Ulviyya Abasova. 2024. *"Decoding Diabetes: Understanding Diagnosis through Medication Prescriptions", International Journal of Recent Advances in Multidisciplinary Research,* 11, (06), xxxx-xxx.

INTRODUCTION

Diabetes is a chronic health condition that affects how the body processes blood glucose (sugar). High levels of blood glucose can lead to serious health problems over time, including heart disease, stroke, kidney disease, nerve damage, and vision loss. Effective management of diabetes is crucial to prevent these complications and maintain quality of life for those affected.

Health Care Costs of Diabetes in the USA: The economic burden of diabetes in the United States is significant and growing. According to the American Diabetes Association (ADA), the total cost of diagnosed diabetes in the U.S. in 2017 was $327 billion. This includes both direct medical costs and indirect costs associated with reduced productivity. There are some factors that influence the diabetes-Race/Ethnicity, Gender, Age, Hospitalization, Medical Procedures and Interventions, Emergency and Inpatient Visits.

Race/Ethnicity: Research has elucidated the role of race and ethnicity in shaping diabetes outcomes and management. Studies have shown that racial and ethnic groups such as Caucasians and Hispanics experience different prevalence rates and complications related to diabetes (American Diabetes Association, 2020). For instance, minority groups, including Hispanics, often have higher rates of diabetes compared to Caucasians, which can be attributed to factors such as genetic predisposition, socioeconomic status, and cultural differences in diet and lifestyle (Golden *et al*.,2012; CDC, 2019). Furthermore, these disparities are exacerbated by limited access to healthcare and educational resources (Kirkman *et al*., 2012).

Gender: Gender differences play a critical role in diabetes outcomes and management. Women with diabetes face unique challenges, such as a higher risk of cardiovascular disease and difficulties in controlling blood glucose levels (Legato *et al*., 2006). Additionally, issues related to pregnancy and menopause can further complicate diabetes management for women (Golden *et al*., 2012). Socioeconomic factors and gender-specific healthcare access also contribute to these disparities, highlighting the need for tailored interventions (ADA, 2020).

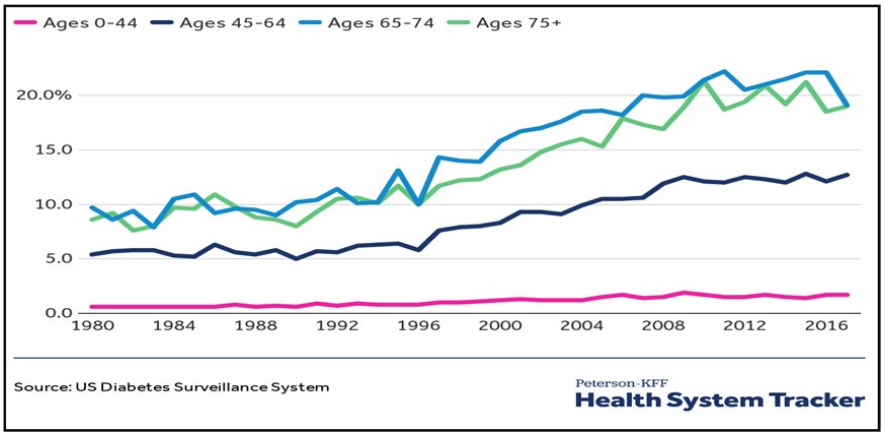


Figure 1.1. Percent of total population with diagnosed diabetes by age, 1980-2017

Age: Age is a crucial factor in the incidence and management of diabetes. Older adults are at a higher risk for type 2 diabetes due to age-related changes in insulin sensitivity and pancreatic function (Sinclair *et al*., 2011). Age also affects the management of diabetes, as older adults often have multiple comorbidities, leading to more complex treatment regimens and a higher likelihood of adverse drug interactions (Kirkman *et al*., 2012). Effective management strategies must therefore consider the unique needs of different age groups (ADA, 2020).

Hospitalization: The length of hospital stay is an important indicator of health outcomes in diabetes patients. Longer hospital stays can be indicative of severe complications or poor management of the disease (Rubin, 2015). Efficient hospital care and early discharge planning are essential to reduce the duration of hospital stays and improve patient outcomes. [1] Studies have highlighted the importance of integrated care approaches to minimize hospital readmissions and manage diabetes more effectively (Nathan *et al*., 2009).

Medical Procedures and Interventions: The number of laboratory procedures and medical interventions a patient undergoes reflects the complexity of their condition and the intensity of their treatment. Frequent lab tests and procedures are often necessary to monitor and manage diabetes, but they also add to the healthcare burden and costs (Zhang *et al*., 2010). Research indicates that regular monitoring and timely medical interventions are critical for effective diabetes management and prevention of complications (Sacks *et al*., 2011).

Emergency and Inpatient Visits: High rates of emergency room visits, and inpatient admissions are associated with poor diabetes control and frequent complications (Ginde *et al*., 2008). Effective outpatient management and patient education can significantly reduce the need for emergency and inpatient care, leading to better health outcomes and lower healthcare costs (Fong *et al*., 2004). Studies underscore the importance of comprehensive diabetes education programs to empower patients in managing their condition and preventing emergencies (Polonsky *et al*., 2011).

The identified research gap underscores the importance of examining the disparity in diabetes diagnosis and treatment across different demographics and healthcare settings. While there is extensive research on diabetes, there is a need for more detailed analysis of how specific factors such as age, gender, ethnicity, and healthcare utilization influence diabetes diagnosis and management.The study investigates the multifaceted factors influencing diabetes diagnosis and prescription patterns in the United States, using data from 130 US hospitals and integrated delivery networks over a ten-year period (1999–2008). The research aims to elucidate the key determinants of diabetes diagnosis and the subsequent prescription of diabetic medication, building on several critical research questions.

Main Research Questions

- How does age influence the likelihood of being prescribed diabetic medication or diagnosed with diabetic
- What is the relationship between race/ethnicity and the likelihood of a diabetes diagnosis?
- How does gender affect the likelihood of a diabetes diagnosis?

Additional Research Questions

- How does the glucose serum level impact the probability of a diabetes diagnosis?
- What is the impact of emergency visits on diabetes diagnosis?
- How do inpatient visits correlate with diabetes diagnosis?

This study aims to fill this gap by analyzing clinical treatment patterns across a diverse set of hospitals and integrated delivery networks in the United States, offering insights into the key determinants of diabetes management.

METHODOLOGY

INFORMATION ABOUT DATA: The dataset includes clinical treatment from 130 US hospitals and integrated delivery networks over a ten-year period (1999–2008). Each row contains hospital records of diabetic patients who were admitted for up to 14 days, underwent testing, and were prescribed medication. Many diabetes patients do not receive numerous preventative and treatment strategies, despite high-quality research suggesting improved clinical outcomes for such individuals. In hospitals, if diabetes isn't managed carefully, it leads to poor control of low blood sugar, which is a problem. Failure to provide proper diabetes care not only increases the managing costs for the hospitals (as the patients are readmitted) but also impacts the morbidity and mortality of the patients, who may face complications associated with diabetes. The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab tests performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc. [2]

DATA PREPROCESSING AND DATA CLEANING: The dataset contains 50 columns and 101766 rows in the dataset.

	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id
1	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1
2	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7
3	64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7
4	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7
5	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7
6	35754	82637451	Caucasian	Male	[50-60)	?	2	1	2
7	55842	84259809	Caucasian	Male	[60-70)	?	3	1	2
8	63768	114882984	Caucasian	Male	[70-80)	?	1	1	7
9	12522	48330783	Caucasian	Female	[80-90)	?	2	1	4

Figure 2.1. A brief overview of the dataset 9x9

During the initial evaluation of the data, it was determined that there were no missing values in the dataset, typically indicated by NA in R. However, a more thorough investigation revealed that missing information was encoded using a special character, the question mark ("?").

This finding underscores the importance of employing comprehensive data cleaning procedures to properly prepare the dataset for econometric analysis. Therefore, to ensure the accuracy and reliability of the analytical results, it is crucial to apply appropriate data cleaning strategies in the presence of these hidden missing values.

Data Cleaning: These columns include “?” signs: "race" , "weight" , "payer code", "medical specialty" , "diag_1", "diag_2", "diag_3". The percentage table of columns which has a high number of special characters is given below.

Table 1.1 The rate of “?” signs in the variables

```
Column: race - Percentage of '?' signs: 2.233555 %
Column: weight - Percentage of '?' signs: 96.85848 %
Column: payer_code - Percentage of '?' signs: 39.55742 %
Column: medical_specialty - Percentage of '?' signs: 49.08221 %
Column: diag_1 - Percentage of '?' signs: 0.02063558 %
Column: diag_2 - Percentage of '?' signs: 0.3517874 %
Column: diag_3 - Percentage of '?' signs: 1.398306 %
```

There are excessive amounts of “?” values in weight, medical specialty, and payer code variables.

The reason why, we eliminated these columns from the dataset. However, we still have missing values in the dataset. As a next step, we removed the rows which contained the “?” sign. The number of columns has declined to 47 from 50. Besides, the 3713 rows deleted, and the number of rows decreased to 98053 from 101766.Rows where gender was labeled as "Unknown/Invalid" were removed. In the readmitted variable, values "<30" and ">30" were changed to "Yes' ' and "No," respectively. The '>200' and '>300' values in the 'max_glu_serum' column were directly recorded to 'High.' The levels of the A1C result variable were also recorded. Using the z-score outlier detection technique, we removed outliers that would negatively impact the model we plan to create. Consequently, 11 rows identified as outliers were deleted from the dataset.Subsequently, the formats of the variables were manually modified. Most of the variables were converted to nominal, with 8 being continuous and only 1 ordinal variable (age).As a final preprocessing step, new columns were created by utilizing elements of the nominal and ordinal variables to aid the model preparation process.

Exploratory Data Analysis: From the plot, it is evident that a significantly larger number of patients are on diabetes medication ("Yes") compared to those who are not ("No"). The bar corresponding to "Yes" is noticeably higher, indicating a frequency greater than 60,000. In contrast, the bar for "No" represents a frequency around 20,000. This distribution suggests that a substantial proportion of the dataset consists of patients who are undergoing diabetes treatment with medication. This insight is crucial for understanding the patient population and could have implications for further analyses, such as assessing the impact of medication on health outcomes or evaluating treatment efficacy. The large number of patients on diabetes medication might also indicate a high prevalence of diabetes within the studied population.

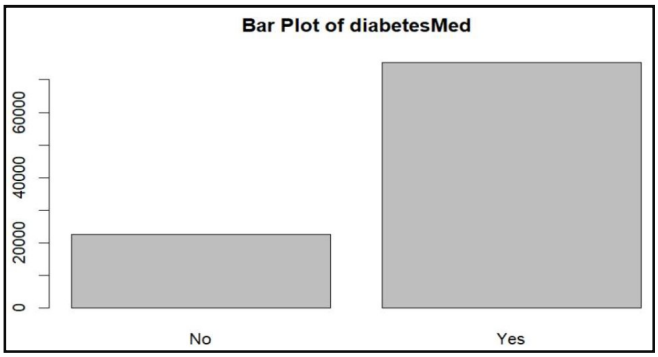


Figure 2.2: Count of the diabetic patients

age <chr>	count <int>
[0-10)	49
[10-20)	392
[20-30)	1163
[30-40)	2677
[40-50)	7073
[50-60)	12970
[60-70)	17073
[70-80)	19552
[80-90)	12507
[90-100)	1884

1-10 of 10 rows

race <chr>	count <int>
AfricanAmerican	14519
Asian	464
Caucasian	57648
Hispanic	1509
Other	1200

5 rows

Figure 2.3. Count of the diabetic patients among different age intervals and distinctive ethnicities

The age distribution suggests a concentration of individuals in the older age categories, particularly between 50 and 80 years. This could imply a focus on older adults within the dataset, which might be relevant for studies on age-related health issues or services targeting older populations.The race distribution shows a predominant representation of Caucasian individuals, followed by African Americans. The relatively small counts for Hispanic, Asian, and other categories indicate a less diverse racial composition in the dataset. This could have implications for analyses related to racial health disparities or the need for more inclusive data collection practices.

Method/Model

Variable selection: In this stage of the selection of the regressors based on correlation, and Antidiabetic medical elements.

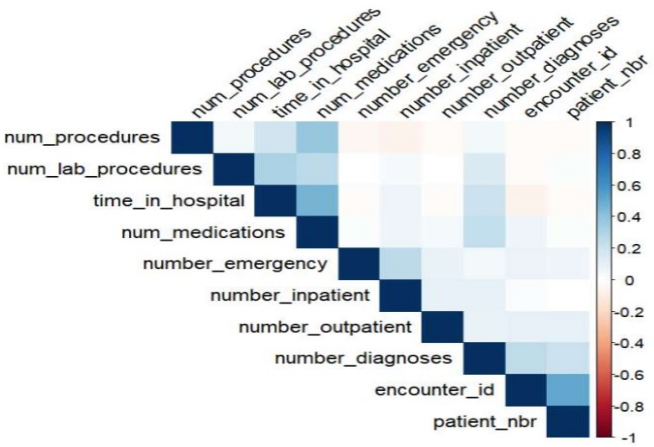


Figure 3.1. The Correlation Matrix plot of the numeric variables

Correlation Analysis: The correlation matrix illustrates the relationships between various variables in the dataset. Notable strong positive correlations are observed between the number of procedures and the number of lab procedures, as well as between the number of inpatient visits and the time spent in the hospital. Additionally, a strong positive correlation exists between the number of diagnoses and the time spent in the hospital. Moderate positive correlations are identified between the number of medications and the number of lab procedures, and between the number of inpatient visits and the number of diagnoses. These relationships suggest that patients with more procedures or diagnoses tend to have more lab tests and longer hospital stays.Weak or negligible correlations are observed among several other variables, indicating limited or no significant relationships.The matrix also highlights potential multicollinearity concerns, particularly between variables such as the number of procedures and lab procedures, and the number of inpatient visits and time in hospital. These high correlations suggest that multicollinearity may pose a problem in regression models, affecting the interpretation of individual predictor effects. Addressing multicollinearity through variable selection or dimensionality reduction techniques will be crucial for accurate model development and analysis. [3]

Feature Selection in Model Development: Through correlation analysis, it was discovered that there is a substantial association between "Patient Number" and "Encounter ID" during the variable selection process. However, neither variable contributed significant explanatory power to the model. Consequently, both variables were excluded from the subsequent analysis. Furthermore, a noteworthy association was identified between the variables "Number of Procedures" and "Number of Medications," as well as between "Number of Medications" and "Time in Hospital."

Due to the potential for multicollinearity to distort regression coefficient estimations and jeopardize the model's statistical validity, the "Number of Medications" variable was removed from the model. Its multicollinearity with two other variables necessitated its exclusion to enhance model performance and reliability. Additionally, the dataset was cleansed of "ID variables" and "diagnosis variables," as these factors were deemed irrelevant to the model's theoretical applicability and prediction accuracy. The following variables are included in the improved dataset for model building after these pointless and troublesome variables were eliminated:

Diabetes Med", "race", "gender", "age", "time_in_hospital", "num_lab_procedures", "num_procedures", "number_outpatient", "number_emergency", "number_inpatient", "number_diagnoses", "max_glu_serum", "A1C result", "change", "readmitted"

Exclusion of Antidiabetic Medications from the Predictive Model: The exclusion of the medical elements refers to their high level of insignificance and proof that is in table below. Anti-diabetic drugs were purposefully left out of the prediction model's development in order to diminish their direct influence on the dependent variable—the existence of diabetes. Making this methodological decision was crucial to improving the model's interpretability and predictive accuracy. Focusing solely on non-pharmacological predictors, the research aims to discover inherent risk factors and biomarkers that indicate the start of diabetes. By avoiding possible confounding effects from therapeutic treatments, this approach increases the robustness of the model and makes it more applicable to predicting the development of diabetes in a diverse patient population.[4]

Table 2.1. The significance levels of the medical elements towards Diabetic result

metforminNo	-3.720e-01	1.053e+03	0.000	0.999718
metforminSteady	1.955e+01	1.066e+03	0.018	0.985366
metforminUp	1.107e-02	1.304e+03	0.000	0.999993
repaglinideNo	-9.670e-01	3.856e+03	0.000	0.999800
repaglinideSteady	1.908e+01	3.902e+03	0.005	0.996098
repaglinideUp	-1.913e-01	4.592e+03	0.000	0.999967
nateglinideNo	-3.096e-01	7.206e+03	0.000	0.999966
nateglinideSteady	1.952e+01	7.255e+03	0.003	0.997853
nateglinideUp	-2.498e-01	8.727e+03	0.000	0.999977
chlorpropamideNo	-2.149e+00	2.923e+04	0.000	0.999941
chlorpropamideSteady	1.931e+01	2.938e+04	0.001	0.999476
chlorpropamideUp	-8.122e-01	3.140e+04	0.000	0.999979
glimepirideNo	-9.120e-01	1.947e+03	0.000	0.999626
glimepirideSteady	1.922e+01	1.974e+03	0.010	0.992232
glimepirideUp	-3.096e-01	2.415e+03	0.000	0.999898
acetohexamideSteady	1.293e+00	2.923e+04	0.000	0.999965
glipizideNo	-8.962e-01	1.124e+03	-0.001	0.999364
glipizideSteady	1.944e+01	1.145e+03	0.017	0.986447
glipizideUp	-5.711e-02	1.462e+03	0.000	0.999969
glyburideNo	-9.091e-01	1.123e+03	-0.001	0.999354
glyburideSteady	1.945e+01	1.147e+03	0.017	0.986476
glyburideUp	7.451e-03	1.453e+03	0.000	0.999996
tolbutamideSteady	2.135e+01	5.364e+03	0.004	0.996824
pioglitazoneNo	-2.733e-01	2.292e+03	0.000	0.999905
pioglitazoneSteady	1.940e+01	2.307e+03	0.008	0.993290
pioglitazoneUp	1.144e-01	2.814e+03	0.000	0.999968
rosiglitazoneNo	-6.539e-01	2.738e+03	0.000	0.999809
rosiglitazoneSteady	1.888e+01	2.752e+03	0.007	0.994525
rosiglitazoneUp	-2.802e-01	3.329e+03	0.000	0.999933
acarboseNo	1.982e+01	1.565e+04	0.001	0.998989
acarboseSteady	3.903e+01	1.569e+04	0.002	0.998015
acarboseUp	1.989e+01	1.757e+04	0.001	0.999097
miglitolNo	-8.903e-01	1.314e+04	0.000	0.999946
miglitolSteady	1.877e+01	1.361e+04	0.001	0.998900
miglitolUp	4.941e-01	2.367e+04	0.000	0.999983
troglitazoneSteady	6.863e-02	1.431e+04	0.000	0.999996
tolazamideSteady	2.095e+01	3.937e+03	0.005	0.995754
tolazamideUp	1.285e+00	2.923e+04	0.000	0.999965

Selection of the Logistic Regression: Selecting between the logit and probit models is the first step in creating the logistic regression model. Applying information criteria measures is essential to distinguish between these two approaches. The fit and complexity of the logit and probit models are compared using metrics such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria balance goodness-of-fit against model simplicity, aiding in the decision-making process. The results from these information criteria measures are used to determine which model is most suitable, ensuring an optimal balance between accuracy and parsimony in the final model selection.

Table 2.2. The significance levels of the medical elements towards Diabetic result

Logit Model:	Probit Model:
AIC: 71586.57	AIC: 71587.12
BIC: 71852.38	BIC: 71852.92

In conclusion, a comparison of the information criteria shows that, in comparison to the Probit model, the Logit model has a lower Akaike Information Criterion (AIC) and a lower Bayesian Information Criterion (BIC). Models with lower values of these criteria generally have better goodness-of-fit and parsimony balance. This analysis concludes that the Logit model works better for our needs than the Probit model. As a result, it has been decided to move forward with the Logit model for more analysis.

Model Preparation: We created a general model using all selected variables and the logit regression model, incorporating two interaction variables. Interaction terms are critical in regression models as they allow for the exploration of combined effects between variables, revealing whether the effect of one variable on the outcome is modified by the presence of another variable. Including these interaction terms can significantly enhance the model's explanatory power, capturing complex relationships that may not be apparent when considering variables independently. This approach ensures a more nuanced understanding of the data, potentially leading to more accurate predictions and better-informed decision-making. [5].

Table 2.3. General Modelwith Interaction Term

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.0735638  0.0649165  16.538  < 2e-16 ***
Caucasian                     -0.3388369  0.0717619  -4.722 2.34e-06 ***
Hispanic                      -0.0636781  0.0557142  -1.143 0.253063
Female                        -0.0717485  0.0154687  -4.638 3.51e-06 ***
ageinterval2                   0.4726333  0.1302403   3.629 0.000285 ***
ageinterval6                   0.1550719  0.0260492   5.953 2.63e-09 ***
ageinterval7                   0.1936938  0.0246434   7.860 3.85e-15 ***
ageinterval8                   0.1379460  0.0238215   5.791 7.00e-09 ***
ageinterval9                  -0.0079204  0.0260398  -0.304 0.761003
time_in_hospital0.0515499  0.0029360  17.558  < 2e-16 ***
num_lab_procedures            -0.0020946  0.0004339  -4.827 1.38e-06 ***
num_procedures                -0.0296137  0.0046062  -6.429 1.28e-10 ***
number_inpatient0.0509673  0.0064894   7.854 4.03e-15 ***
number_diagnoses              -0.0172474  0.0082836  -2.082 0.037332 *
max_glu_serum_Norm            -0.5640451  0.0433120 -13.023  < 2e-16 ***
A1Cresult_High                 0.3623036  0.0952013   3.806 0.000141 ***
Caucasian:number_diagnoses     0.0467101  0.0094940   4.920 8.66e-07 ***
num_lab_procedures:A1Cresult_High  0.0076018  0.0017253   4.406 1.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two interaction terms have been created to explain model well.

Caucasian (Estimate: 0.0467101, p-value: 8.66e-07):This interaction term highlights how the effect of the number of diagnoses on the outcome differs for Caucasian patients compared to other racial groups. The positive coefficient suggests that as the number of diagnoses increases, the impact on the outcome is more pronounced for Caucasian patients.

Num_lab_procedures (Estimate: 0.0076018, p-value: 1.05e-05): This interaction term shows the combined effect of the number of lab procedures and a high A1C result on the outcome. The positive coefficient indicates that having a high number of lab procedures combined with a high A1C result has a significant and positive impact on the outcome. Including these interaction terms is crucial as they capture the combined effects of variables, providing a more nuanced understanding of the relationships in our data. This enhances the explanatory power of our model and helps in making more accurate predictions. Using the general-to-specific approach, we initially examined whether all insignificant variables could be removed from the model simultaneously. Our hypothesis testing framework was as follows:

H0: The insignificant variables are jointly insignificant
H1: The insignificant variables are jointly significant

Given the p-value of 2.2e-16 ***, we rejected the null hypothesis. This indicates that the insignificant variables are jointly significant in our model, precluding the simultaneous elimination of all these variables. Consequently, we proceeded to drop variables one by one based on their p-values, systematically removing the most insignificant variable from the general regression model. Through this iterative process, we eliminated all insignificant variables, except for the "change_ch" variable, which remained in the model due to its potential relevance. This methodical approach ensured that our final model retained only the variables that contribute significantly to the explanation of the dependent variable, enhancing both the robustness and the interpretability of the model. As a result, we now have a refined and ready model, optimized for accuracy and reliability in its predictive capabilities.

Final Model and Model Interpretation

Final Model

Table 2.4. Final Logistic Regression Model

```
Coefficients:

                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      0.4609167  0.0750025   6.145 7.98e-10 ***
Caucasian                       -0.3191247  0.0824726  -3.869 0.000109 ***
Hispanic                        -0.1821721  0.0661823  -2.753 0.005913 **
Female                          -0.0505185  0.0179060  -2.821 0.004783 **
ageinterval2                     0.6984450  0.1403786   4.975 6.51e-07 ***
ageinterval6                     0.1573285  0.0305031   5.158 2.50e-07 ***
ageinterval7                     0.2058093  0.0288423   7.136 9.63e-13 ***
ageinterval8                     0.2358786  0.0277629   8.496  < 2e-16 ***
ageinterval9                     0.1543589  0.0302967   5.095 3.49e-07 ***
time_in_hospital0.0111560  0.0033399   3.340 0.000837 ***
num_lab_procedures              -0.0024448  0.0005052  -4.839 1.30e-06 ***
num_procedures                  -0.0170469  0.0053959  -3.159 0.001582 **
number_emergency0.0266251  0.0122396   2.175 0.029606 *
number_inpatient0.0424048  0.0076553   5.539 3.04e-08 ***
number_diagnoses                -0.0369729  0.0095567  -3.869 0.000109 ***
max_glu_serum_Norm              -0.3730312  0.0498290  -7.486 7.09e-14 ***
A1Cresult_High                   0.2296140  0.1077890   2.130 0.033154 *
change_Ch                       19.2513089 50.3733744   0.382 0.702334
Caucasian:number_diagnoses       0.0402042  0.0109624   3.667 0.000245 ***
num_lab_procedures:A1Cresult_High  0.0045020  0.0019559   2.302 0.021347 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 106110  on 98041  degrees of freedom
Residual deviance:  71690  on 98022  degrees of freedom
AIC: 71730

Number of Fisher Scoring iterations: 18
```

In the logit model, we cannot interpret our variables quantitatively due to the nature of logistic regression. The coefficients in a logistic regression model represent the log odds of the dependent variable occurring for a one-unit change in the predictor variable, not the direct change in probability. This makes the interpretation less intuitive compared to linear regression coefficients, which represent the change in the dependent variable directly. Before interpreting the parameters qualitatively in this logistic regression model, we first examine the p-values to determine whether each variable is significant or insignificant. If a variable is insignificant, it has no effect on the dependent variable, and we do not interpret this regressor. For this reason, we will not interpret the "change_Ch" variable, as it has no impact. During the interpretation process, we also consider the base levels, which are not explicitly given in the model output. The base levels serve as reference categories against which other categories are compared. In econometrics, a value of 1 is considered a success; this does not carry any linguistic meaning but rather a statistical convention. In the context of our output, "success" likely refers to the event of being diabetic. This interpretation aligns with the typical convention in logistic regression, where one outcome (e.g., being diabetic) is considered the "success" or "positive" outcome, while the other outcome (e.g., not being diabetic) is considered the "failure" or "negative" outcome. This qualitative approach allows us to understand the direction and relative importance of the predictors in relation to the dependent variable without misinterpreting the coefficients' magnitude. By focusing on the significance and direction of the effects, we can derive meaningful insights from the model while adhering to the appropriate interpretation guidelines for logistic regression.

Interpretation of the Variables from the Ready Model

Caucasian: Being Caucasian is associated with a decreased probability of being diabetic compared to other races.

Hispanic: The likelihood of being diabetic is lower for Hispanic individuals compared to other races.
Female: Females have a reduced probability of being diabetic compared to males.

Age Interval 2 (10-20 years): Individuals aged 10-20 have a higher probability of being diabetic compared to those aged 90-100.

Age Interval 6 (50-60 years): The probability of being diabetic is higher for individuals aged 50-60 compared to those aged 90-100.

Age Interval 7 (60-70 years): Being in the 60-70 age range increases the likelihood of being diabetic compared to the 90-100 age range.

Age Interval 8 (70-80 years): Individuals aged 70-80 are more likely to be diabetic than those aged 90-100.

Age Interval 9 (80-90 years): The likelihood of being diabetic is higher for those aged 80- 90 compared to individuals aged 90-100.

Time in Hospital: Spending more days in the hospital increases the probability of being diabetic.

Number of Lab Procedures: A higher number of lab procedures is associated with a decreased probability of being diabetic.

Number of Procedures: An increased number of medical procedures reduces the likelihood of being diabetic.

Number of Emergency Visits: More emergency visits correlate with a higher probability of being diabetic.

Number of Inpatient Visits: A greater number of inpatient visits increases the likelihood of being diabetic.

Number of Diagnoses: Having more diagnoses is associated with a decreased probability of being diabetic.

Normal Max Glucose Serum: Normal maximum glucose serum levels decrease the likelihood of being diabetic compared to high glucose levels.

High A1C Result: A high A1C result increases the probability of being diabetic compared to a normal result.

Change in Medications (change_Ch): Changes in medications do not significantly affect the probability of being diabetic and are not interpreted further.

Interaction: Caucasian and Number of Diagnoses: For Caucasian individuals, a higher number of diagnoses increases the probability of being diabetic.

Interaction: Number of Lab Procedures and High A1C Result: The combined effect of more lab procedures and a high A1C result increases the likelihood of being diabetic.

Model Evaluation / Diagnostic Tests: Diagnostic tests are crucial for assessing the validity and goodness-of-fit of logistic regression models. These tests help identify model specification errors, evaluate the model's predictive accuracy, and ensure the robustness of the results. In this analysis, we employed the Linktest and the Hosmer-Lemeshow test to evaluate our logistic regression model.

Linktest Diagnostic Test: The Linktest is used to detect model specification errors by testing whether the model is correctly specified.

Table 2.5. Coefficients of the Linktest result

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.001598   0.016417  -0.097    0.922
yhat         1.011394   0.051817  19.519<2e-16 ***
yhat2       -0.014237   0.016121  -0.883    0.377
```

It does this by adding the predicted values (yhat) and the squared predicted values (yhat2) as predictors in a new regression model. The null hypothesis (H0) is that the model is correctly specified.

Based on our results

Coefficient for yhat is 1.011392, p value <2e-16 (close to zero). Coefficient for yhat2 is -0.014237, p value 0.377. Yhat is significant and yhat2 is insignificant,suggest that we fail to reject the null hypothesis. This indicates that the model specifications are appropriate, and there are no significant specification errors.

Hosmer Lemeshow Diagnostics test: The Hosmer-Lemeshow test evaluates the goodness-of-fit of the logistic regression model. It compares the observed and expected frequencies in different groups of the data, testing the null hypothesis (H0) that the model fits the data well

Table 2.6. Coefficients of theHosmer Lemeshow Diagnostics test result

```
        Hosmer and Lemeshow test (binary model)

data:  dataset_selected$diabetesMed, fitted(reg10)
   X-squared = 1.479, df = 8, p-value = 0.9931
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that the model fits the data well. The high p-value suggests that the model's predictions are consistent with the observed outcomes, confirming the adequacy of the model.

Other Diagnostic Tests (Not Performed in This Analysis): While we did not perform the Osius-Rojek test and the Stukel test in this analysis, they are also valuable for diagnosing logistic regression models:

- **Osius-Rojek Test:** This test is used to detect general lack of fit in logistic regression models. It evaluates whether the model adequately fits the data by comparing observed and predicted values.
- **Stukel Test:** This test is designed to detect deviations from linearity in the logit model. It assesses whether the logistic regression assumption of linearity in the logit is violated.

In this analysis, the Linktest and Hosmer-Lemeshow test were employed to validate our logistic regression model. The results from both tests indicated that the model is correctly specified and fits the data well, thus confirming the robustness and reliability of our model specifications. These diagnostic tests are essential for ensuring the integrity of econometric analyses and the validity of the resulting inferences.

Model Comparison with Restricted Models: The model comparison is implemented via using ready model versus three restricted models. For this comparison the Likelihood Ratio test has been utilized.

Information about Likelihood Ratio Test: In econometric analysis, the likelihood ratio test is used to compare the goodness-of-fit between two models: a restricted model and an unrestricted model. The unrestricted model includes all potential explanatory variables, while the restricted model excludes one or more of these variables. The hypothesis tested is:

H0: The restricted model is adequate, meaning the excluded variables do not significantly improve the model.

H1: The unrestricted model is better, meaning the excluded variables do significantly improve

Results of the Likelihood Ratio Test

We compare the unrestricted model with three restricted models. The unrestricted model (Main Model) includes all selected variables, while the restricted models exclude certain variables to test their joint significance.

Unrestricted Model (Main Model)

- **Log Likelihood**: -35,844.760
- **Akaike Information Criterion (AIC)**: 71,729.510

Restricted Model 1 (Rest 1)

- **Log Likelihood**: -53,055.140
- **AIC**: 106,112.300

Restricted Model 2 (Rest 2)

- **Log Likelihood**: -52,836.260
- **AIC**: 105,682.500

Restricted Model 3 (Rest 3)

- **Log Likelihood**: -52,924.870

Calculation of the Likelihood Ratio Test Statistic

1. **Restricted Model 1 vs. Unrestricted Model:**
   $\lambda_1 = -2(-53,055.140 + 35,844.760) = 34,420.760$
   Degrees of Freedom (df): Number of restrictions (variables excluded) from the unrestricted model.

2. **Restricted Model 2 vs. Unrestricted Model:**
   $\lambda_2 = -2(-52,836.260 + 35,844.760) = 33,983.000$
   Degrees of Freedom (df): Number of restrictions (variables excluded) from the unrestricted model.

3. **Restricted Model 3 vs. Unrestricted Model:**
   $\lambda_3 = -2(-52,924.870 + 35,844.760) = 34,160.220$
   Degrees of Freedom (df): Number of restrictions (variables excluded) from the unrestricted model.

Figure 3.3. The calculation of the Likelihood Ratio Test Statistic

Interpretation of Results: The calculated λ values for all three comparisons are significantly high. Given the large sample size (98,042 observations) and the chi-squared distribution of the test statistic, we compare these values against the critical values from the chi-squared distribution table. Typically, with such high values, the null hypothesis (H0) is rejected. For Restricted Model 1, the very high λ1value suggests that the excluded variables significantly improve the model. Hence, the restricted model is not adequate. For Restricted Model 2, the λ2 value also indicates that the excluded variables are important, rejecting the adequacy of the restricted model. For Restricted Model 3, the λ3 value again shows that the restricted model is inadequate, as the excluded variables significantly enhance the model.

Based on the likelihood ratio tests, we reject the null hypothesis for all three restricted models. This indicates that the excluded variables in each case are jointly significant and contribute meaningfully to the model. Therefore, the unrestricted model, which includes all variables, provides the best fit and should be preferred for our econometric analysis. The robustness of the unrestricted model is further supported by its lower AIC value compared to the restricted models, indicating better overall model performance.

Marginal Effects and interpretation

In econometric analysis, particularly in logistic regression models, interpreting the marginal effects provides insights into the impact of each predictor variable on the probability of the outcome variable. Marginal effects can be calculated in two primary ways: Marginal Effects at the Means (MEM) and Average Marginal Effects (AME). Using these two approaches ensures that we capture both the typical and average impacts of our predictors, enhancing the robustness and interpretability of our econometric analysis. The results from the atmean = FALSE Average Marginal Effects (AME) model appear to be more reliable and informative, demonstrating significant relationships for multiple variables. In contrast, the atmean = TRUE Marginal Effects at the Means (MEM) model indicates almost no significant effects, except forchange_Ch, which might not accurately reflect the true relationships within the data.Therefore, based on these results, it is advisable to consider the findings from the atmean = FALSE model, as they seem to provide more meaningful insights into the effects of the predictor variables on the probability of the outcome. Consequently, we will not interpret the results from the atmean = TRUE model, as they do not offer as robust or valid conclusions regarding the predictor variables' impact. The following interpretation of marginal effects is based on the Average Marginal Effects (AME) model, which provides reliable and informative insights into the effects of predictor variables on the probability of being diabetic.

Table 2.7. Comparison of Logistic Regression Models

Comparison of Logistic Regression Models

	Dependent Variable: diabetesMed			
	diabetesMed			
	Main Model (1)	Restricted Model 1 (2)	Restricted Model 2 (3)	Restricted Model 3 (4)
Caucasian	-0.319***		-0.001	
	(0.082)		(0.018)	
Hispanic	-0.182***			
	(0.066)			
Female	-0.051***			
	(0.018)			
ageinterval2	0.698***			
	(0.140)			
ageinterval6	0.157***		0.082***	
	(0.031)		(0.021)	
ageinterval7	0.206***			
	(0.029)			
ageinterval8	0.236***			
	(0.028)			
ageinterval9	0.154***			
	(0.030)			
time_in_hospital	0.011***		0.055***	
	(0.003)		(0.003)	
num_lab_procedures	-0.002***			0.003***
	(0.001)			(0.0004)
num_procedures	-0.017***		-0.025***	-0.007
	(0.005)		(0.004)	(0.004)
number_emergency	0.027**			0.080***
	(0.012)			(0.012)
number_inpatient	0.042***			0.033***
	(0.008)			(0.007)
number_diagnoses	-0.037***			0.025***
	(0.010)			(0.004)
max_glu_serum_Norm	-0.373***			
	(0.050)			
A1Cresult_High	0.230**			
	(0.108)			
change_Ch	19.251			
	(50.373)			
Caucasian:number_diagnoses	0.040***			
	(0.011)			
num_lab_procedures:A1Cresult_High	0.005**			
	(0.002)			
Constant	0.461***	1.200***	0.984***	0.842***
	(0.075)	(0.008)	(0.020)	(0.034)
Observations	98,042	98,042	98,042	98,042
Log Likelihood	-35,844.760	-53,055.140	-52,836.260	-52,924.870
Akaike Inf. Crit.	71,729.510	106,112.300	105,682.500	105,861.700
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2.8. Marginal Effects of the regressors

Marginal Effects:

	dF/dx	Std. Err.	z	P>|z|	
Caucasian	-4.0978e-02	1.0308e-02	-3.9752	7.032e-05	***
Hispanic	-2.4070e-02	8.8051e-03	-2.7337	0.0062627	**
Female	-6.6099e-03	2.3419e-03	-2.8225	0.0047653	**
ageinterval2	8.4177e-02	1.4975e-02	5.6213	1.895e-08	***
ageinterval6	2.0403e-02	3.9124e-03	5.2150	1.838e-07	***
ageinterval7	2.6655e-02	3.6844e-03	7.2344	4.677e-13	***
ageinterval8	3.0559e-02	3.5455e-03	8.6193	< 2.2e-16	***
ageinterval9	2.0023e-02	3.8883e-03	5.1496	2.610e-07	***
time_in_hospital	1.4600e-03	4.3728e-04	3.3387	0.0008416	***
num_lab_procedures	-3.1995e-04	6.6179e-05	-4.8346	1.334e-06	***
num_procedures	-2.2309e-03	7.0644e-04	-3.1580	0.0015887	**
number_emergency	3.4844e-03	1.6021e-03	2.1749	0.0296339	*
number_inpatient	5.5495e-03	1.0030e-03	5.5327	3.153e-08	***
number_diagnoses	-4.8386e-03	1.2514e-03	-3.8666	0.0001104	***
max_glu_serum_Norm	-4.9540e-02	6.6396e-03	-7.4613	8.566e-14	***
A1Cresult_High	2.9631e-02	1.3659e-02	2.1693	0.0300612	*
change_Ch	4.2533e-01	2.1698e-03	196.0197	< 2.2e-16	***
Caucasian:number_diagnoses	5.2615e-03	1.4354e-03	3.6656	0.0002468	***
num_lab_procedures:A1Cresult_High	5.8918e-04	2.5601e-04	2.3013	0.0213727	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model Performance Metrics: To evaluate the performance of our logistic regression model, we focus on three key metrics: Count R-squared, Adjusted Count R-squared, and McKelvey &Zavoina R-squared. These metrics provide insights into the model's predictive accuracy and its ability to explain the variance in the outcome variable.

Table 2.9. Model Performance Metrics

McFadden	Adj.McFadden	Cox.Snell	Nagelkerke	McKelvey.Zavoina
3.243867e-01	3.239909e-01	2.960738e-01	4.477943e-01	9.656825e-01

Effron	Count	Adj.Count	AIC	Corrected.AIC
2.645401e-01	7.697619e-01	5.682319e-03	7.172951e+04	7.172952e+04

Key metrics

Count R-squared

- **Value:** 0.7698
- **Interpretation:** The Count R-squared measures the proportion of correct predictions made by the model. A value of 0.7698 indicates that the model correctly predicts approximately 76.98% of the cases. This high percentage suggests that the model has good predictive accuracy and is effective in correctly classifying the outcome. In other words, the model correctly predicts 77% of all observations.

Adjusted Count R-squared

- **Value:** 0.0057
- **Interpretation:** The Adjusted Count R-squared adjusts the Count R-squared for chance agreement. A value of 0.0057 indicates a slight improvement over random guessing. This lower value compared to the Count R-squared suggests that while the model performs well in prediction, the adjustment for random chance provides a more conservative measure of the model's performance. Essentially, 0.57% of all predictions were correct solely due to the regressors, accounting for the variation of the dependent variable.

McKelvey & Zavoina R-squared

- **Value:** 0.9657
- **Interpretation:** The McKelvey &Zavoina R-squared is designed to approximate the R-squared from linear regression models. A value of 0.9657 indicates an excellent fit, suggesting that the model explains a substantial portion of the variance in the outcome variable. This high value highlights the model's effectiveness in capturing the underlying patterns in the data. If the latent (unobserved) variable were observed, our model would explain 97% of its variation.

Additional Metrics

- **McFadden R-squared:** Although not directly interpretable, it is used for comparing model fit among different models.
- **Adjusted McFadden R-squared:** An adjustment of the McFadden R-squared for the number of predictors in the model.
- **Cox & Snell R-squared:** Another measure of the model's explanatory power.
- **Nagelkerke R-squared:** A version of the Cox & Snell R-squared that adjusts the scale to cover the full range from 0 to 1.
- **Akaike Information Criterion (AIC):** A measure of model fit that penalizes for the number of parameters; lower values indicate a better fit.

The model performance metrics indicate that the logistic regression model has good predictive accuracy and effectively explains the variance in the outcome variable. The Count R-squared and McKelvey &Zavoina R-squared values demonstrate the model's strength in prediction and variance explanation, respectively. However, the Adjusted Count R-squared suggests a need for cautious interpretation, accounting for the potential influence of random chance. Overall, these metrics provide a comprehensive evaluation of the model's performance, supporting its robustness and reliability in predicting the probability of the outcome variable.

FINDINGS

Our research aimed to explore the factors influencing the likelihood of being prescribed diabetic medication or diagnosed with diabetes. Through a comprehensive econometric analysis using logistic regression, we addressed both main and additional research questions.

Main Research Questions

- **How does age influence the likelihood of being prescribed diabetic medication or diagnosedwith diabetes?**

The analysis indicates that older age groups have an increased probability of being diagnosed with diabetes. Specifically, the age intervals for 50-60 (ageinterval6), 60-70 (ageinterval7), 70-80 (ageinterval8), and 80-90 (ageinterval9) all have positive coefficients, suggesting that older patients, particularly those aged 60-70 and above, are more likely to be diagnosed with diabetes.

- **What is the relationship between race/ethnicity and the likelihood of a diabetes diagnosis?**

The variable for Caucasian patients has a negative coefficient, indicating that Caucasian individuals are less likely to be diagnosed with diabetes compared to other ethnic groups. This finding suggests that race/ethnicity plays a significant role in the likelihood of a diabetes diagnosis.

- **How does gender affect the likelihood of a diabetes diagnosis?**

The analysis shows that the variable for Female has a negative coefficient, suggesting that being female decreases the probability of being diagnosed with diabetes compared to being male. This implies that gender is a significant factor, with female patients having a lower likelihood of receiving a diabetes diagnosis.

Additional Research Questions

How does the glucose serum level impact the probability of a diabetes diagnosis?: Analysis and Conclusion: The variable representing normal glucose serum levels (max_glu_serum_Norm) has a negative coefficient. This indicates that having a normal glucose serum level decreases the probability of being diagnosed with diabetes compared to having an abnormal glucose serum level. Thus, the glucose serum level is a significant factor in diabetes diagnosis, with normal levels being associated with a lower likelihood of receiving a diabetes diagnosis. This finding underscores the importance of monitoring glucose levels as part of diabetes management and prevention strategies.

What is the impact of emergency visits on diabetes diagnosis?: Analysis and Conclusion: The variable for the number of emergency visits (number_emergency) has a positive coefficient. This suggests that an increase in the number of emergency visits in the year preceding the encounter is associated with a higher probability of being diagnosed with diabetes. This indicates that frequent emergency healthcare utilization may be an indicator of underlying health issues, including diabetes. Consequently, patients with multiple emergency visits should be closely monitored for diabetes and other chronic conditions, allowing for timely interventions and improved health outcomes.

How do inpatient visits correlate with diabetes diagnosis?

Analysis and Conclusion: The number of inpatient visits (number_inpatient) also has a positive coefficient. This implies that an increase in the number of inpatient visits in the year preceding the encounter correlates with a higher probability of being diagnosed with diabetes. This relationship highlights the connection between intensive healthcare utilization and diabetes diagnosis. Patients with frequent inpatient admissions may have complex health needs, including a higher risk of diabetes, necessitating comprehensive care approaches to address multiple health concerns effectively**.**

In summary, the additional research questions reveal important insights into the factors influencing the probability of a diabetes diagnosis. Normal glucose serum levels are associated with a lower likelihood of diabetes, while frequent emergency and inpatient visits are linked to a higher probability of being diagnosed with diabetes. These findings emphasize the need for continuous monitoring and comprehensive care for patients with frequent healthcare interactions, to ensure early detection and effective management of diabetes.

CONCLUSION

Through a thorough econometric analysis employing logistic regression, this study sought to investigate the factors impacting the likelihood of receiving a diabetes diagnosis or prescription for diabetic medication. The results offer insightful information about how different clinical and demographic factors affect diabetes diagnosis.

Main Findings

- **Age:** Older age groups are significantly more likely to be diagnosed with diabetes. The analysis reveals positive coefficients for age intervals 50-60, 60-70, 70-80, and 80-90, indicating that the likelihood of diabetes diagnosis increases with age.

- **Race/Ethnicity:** The analysis shows that Caucasian patients have a lower probability of being diagnosed with diabetes compared to other ethnic groups, as evidenced by the negative coefficient for the variable Caucasian.
- **Gender:** Female patients are less likely to be diagnosed with diabetes compared to male patients. The variable Female has a negative coefficient, indicating a lower probability of diabetes diagnosis for females.
- **Glucose Serum Levels:** Patients with normal glucose serum levels have a significantly lower probability of being diagnosed with diabetes compared to those with abnormal glucose serum levels. This underscores the importance of maintaining normal glucose levels for diabetes prevention.
- **Emergency Visits:** An increase in the number of emergency visits in the year preceding the encounter is associated with a higher probability of being diagnosed with diabetes. This suggests that frequent emergency healthcare utilization may indicate underlying health issues, including diabetes.
- **Inpatient Visits:** Similarly, a higher number of inpatient visits in the year preceding the encounter correlates with an increased likelihood of diabetes diagnosis. This highlights the need for comprehensive care approaches for patients with frequent inpatient admissions.
- **Interaction Effects:** The study also identified significant interaction effects. For instance, the interaction between being Caucasian and the number of diagnoses increases the probability of diabetes diagnosis. Additionally, the interaction between the number of lab procedures and a high A1C result significantly increases the likelihood of being diagnosed with diabetes.

Implications

The results of this study have important implications for healthcare policies and clinical practices:

- **Age-specific Interventions:** The increased probability of diabetes diagnosis in older age groups suggests the need for age-specific interventions and monitoring to manage and prevent diabetes effectively.
- **Ethnic Disparities:** The lower likelihood of diabetes diagnosis among Caucasians compared to other ethnic groups highlights the importance of addressing ethnic disparities in diabetes care and management.
- **Gender-specific Strategies:** The lower probability of diabetes diagnosis among females indicates the need for gender-specific strategies in diabetes prevention and management.
- **Monitoring Glucose Levels:** Maintaining normal glucose serum levels is crucial for reducing the risk of diabetes, emphasizing the importance of regular monitoring and management of glucose levels.
- **Emergency and Inpatient Care:** Frequent emergency and inpatient visits should trigger closer monitoring for diabetes, allowing for timely interventions and improved health outcomes.
- Comprehensive Care Approaches: The significant interaction effects indicate the necessity of considering combined effects of different variables, leading to more accurate predictions and better-informed decision-making in diabetes care.

In conclusion, the findings from this study contribute to a deeper understanding of the factors associated with diabetes diagnosis. These insights can inform targeted interventions, improve clinical practices, and shape healthcare policies aimed at enhancing diabetes management and prevention efforts.

BIBLIOGRAPHY

1. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed Res Int. 2014;2014:781670. doi: 10.1155/2014/781670. Epub 2014 Apr 3. PMID: 24804245; PMCID: PMC3996476.
2. Clore,John, Cios,Krzysztof, DeShazo,Jon, and Strack,Beata. (2014). Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J.
3. Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.
4. Shrestha, Noora. (2020). Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics. 8. 39-42. 10.12691/ajams-8-2-1.
5. American Diabetes Association. (2020). Standards of Medical Care in Diabetes—2020
6. Centers for Disease Control and Prevention. (2019). National Diabetes Statistics Report.
7. Golden, S. H., *et al*. (2012). Diabetes and women's health.
8. Legato, M. J., *et al*. (2006). Gender-specific care for diabetes.
9. Kirkman, M. S., *et al*. (2012). Diabetes in older adults.
10. Sinclair, A. J., *et al*. (2011). Diabetes in the elderly.
11. Rubin, D. J. (2015). Hospital readmission of patients with diabetes.
12. Ginde, A. A., *et al*. (2008). Trends in emergency department visits for hypoglycemia.
13. Nathan, D. M., *et al*. (2009). Medical management of hyperglycemia in type 2 diabetes.
14. Sacks, D. B., *et al*. (2011). Guidelines and recommendations for laboratory analysis in diabetes.
15. Polonsky, W. H., *et al*. (2011). Assessing psychosocial distress in diabetes
16. Fong, D. S., *et al*. (2004). Retinopathy in diabetes.
17. Zhang, P., *et al*. (2010). The economic impact of diabetes.

9.APPENDIX

Creators of the dataset: John Clore , Krzysztof Cios, Jon DeShazo , Beata Strack

Details about citations of the data: 1 citations and 60346 views Information was extracted from the database for encounters that satisfied the following criteria. (1) It is an inpatient encounter (a hospital admission). (2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered into the system as a diagnosis. (3) The length of stay was at least 1 day and at most 14 days. (4) Laboratory tests were performed during the encounter. (5) Medications were administered during the encounter.

There are 50 columns and 101766 rows in the dataset. Information about the variables:

Encounter ID: Unique identifier for each visit.
Patient Number (Patient_nbr): Unique identifier for each patient.
Race: Racial background of the patient.
Gender: Gender of the patient.
Age: Age group of the patient.
Weight: Weight of the patient (if available).
Admission Type ID: Type of admission (e.g., emergency, elective).
Discharge Disposition ID: Type of discharge (e.g., home, another facility).
Admission Source ID: Source of admission (e.g., ER, physician referral).
Time in Hospital: Number of days spent in the hospital.
Payer Code: Code of the responsible healthcare payer.
Medical Specialty: Specialty of the treating physician.
Number of Lab Procedures: Count of laboratory procedures performed.
Number of Non-Lab Procedures: Count of non-laboratory procedures.
Number of Medications: Count of distinct medications administered.
Number of Outpatient Visits: Number of outpatient visits in the past year.
Number of Emergency Visits: Number of emergency visits in the past year.
Number of Inpatient Visits: Number of inpatient visits in the past year.
Diagnosis Codes (Diag_1, Diag_2, Diag_3): Primary, secondary, and tertiary diagnosis.
Number of Diagnoses: Total number of diagnoses recorded.
Max Glucose Serum Level: Result of latest blood glucose level test.
A1C Test Result: Result of latest A1C test.
Medications: Indicators for prescribed medications.
Medication Combinations: Indicators for specific medication combinations.
Change in Medication: Indicator for change in medication during encounter.
Diabetic Medication Prescribed: Indicator for diabetic medication prescription.
Readmission: Indicator for readmission within 30 days after encounter.

Data Types

Ordinal Variables: 1. age

Nominal Variables: 2. race, 3. gender, 4. max_glu_serum, 5. A1Cresult , 6. Metformin , 7. repaglinide , 8. nateglinide , 9. chlorpropamide , 10. glimepiride, 11. acetohexamide, 12. glipizide, 13. glyburide, 14. tolbutamide, 15. pioglitazone, 16. rosiglitazone, 17. acarbose, 18. miglitol, 19. troglitazone, 20. tolazamide, 21. examide, 22. citoglipton, 23. insulin, 24. glyburide.metformin, 25. glipizide.metformin, 26. glimepiride.pioglitazone, 27. metformin.rosiglitazone, 28. metformin.pioglitazone, 29. change, 30. diabetesMed, 31. Readmitted,32. Diog1, 33. Diog2, 34. Diog3, 35. discharge_disposition_id, 36. admission_source_id, 37. dataset$admission_type_id

Numerical Variables: 38. time_in_hospital, 39. num_lab_procedures, 40. num_procedures, 41. num_medications, 42. number_outpatient,43number_emergency,44number_inpatient,45.number_diagnoses

Non-measurable **variables:** 46. encounter_id, 47. patient_nbr

NOTE: The diagnostic tests: Osius Rojek and Stukel have not been used because of the issue in the coding part.

ESTIMATOR OF THE DIABETIC APP

A Diabetic Calculator app where users may estimate the chance of being diabetic in the near future based on an ADA test which is a reliable https://turgudvaliyev.shinyapps.io/Diabetic-Calculator/

To estimate a person's chance of being diabetic by percentage, several factors need to be considered. These factors include:

- **Age:** The risk of diabetes increases with age.
- **Family History:** A family history of diabetes significantly raises the risk.
- **Body Mass Index (BMI):** Higher BMI is a strong risk factor for diabetes.
- **Waist Circumference:** Abdominal obesity is closely linked to diabetes risk.
- **Physical Activity Level:** Sedentary lifestyle increases the risk.
- **Diet:** A diet high in sugars and fats can raise the risk of diabetes.
- **Blood Pressure:** Hypertension is often associated with diabetes.
- **Cholesterol Levels:** Dyslipidemia (abnormal cholesterol levels) can be a risk factor.
- **History of Gestational Diabetes:** Women who had diabetes during pregnancy are at higher risk.
- **Ethnicity:** Certain ethnic groups (e.g., African American, Hispanic, Native American, Asian) have higher risks.
- **Medical History:** Conditions like polycystic ovary syndrome (PCOS) and prediabetes can increase the risk.

Example Calculation Using the ADA Diabetes Risk Test:The American Diabetes Association (ADA) provides a risk test that scores these factors. Here's a simplified version for illustrative purposes:

Age:

- <40 years: 0 points
- 40-49 years: 1 point
- 50-59 years: 2 points
- ≥60 years: 3 points

BMI:

- <25: 0 points
- 25-29: 1 point
- ≥30: 2 points

Family History:

- No: 0 points
- Yes: 1 point

Physical Activity:

- Yes: 0 points
- No: 1 point

History of Hypertension:

- No: 0 points
- Yes: 1 point

History of Gestational Diabetes:

- No: 0 points
- Yes: 1 point

Scoring:

- Low risk: 0-2 points
- Moderate risk: 3-5 points
- High risk: 6-8 points

For instance, a 55-year-old (2 points) with a BMI of 32 (2 points), who has a family history of diabetes (1 point), and is physically inactive (1 point), would score 6 points, indicating a high risk of diabetes. To provide a precise percentage, these factors can be plugged into more detailed risk calculators or models that use population data. Generally, a person in the high-risk category might have a risk percentage significantly above the general population average, which is around 10-15% for adults in many developed countries. For a personalized assessment, consulting with a healthcare provider is recommended. They may use tools like the Framingham Risk Score, UKPDS Risk Engine, or ADA Risk Test to give a more accurate percentage based on comprehensive medical evaluation.

****