Analyzing effectiveness of Metformin across various age groups and ethnicities

Abou Keita, Charishma Tummala, Jaydeep Radadia

December 2021

1 Introduction

According to a Centres for Diseases Control and Prevention report, Type 2 diabetes has been found to be more common than type 1 diabetes. The number of people diagnosed with type 2 diabetes are six times the number of type 1 diabetes [1]. Metformin is considered as a standard when it comes to the initial treatment. If the condition of the patient worsens, then he is subjected to further treatments. The drug was approved by the FDA in 1995. For this reason, we chose to work with Metaformin.

2 Overview of the Dataset

The dataset was obtained from the University of California Irvine's Machine Learning Repository. The data was recorded by the Cerner Corporation, an IT company working in the field of health services. This dataset contains the recorded data between 1998-2008. The dataset has been divided into three tables namely; the main table, the patient id table and the method of admission and discharge table. In the main table, the encounter id serves as a unique identifier. Each time a patient is admitted into the hospital a unique case ID is assigned to the patient. At the same time patient's data is also recorded including of the age, gender, ethnicity, weight, and the mode of admission. In the patient id table, the patient id serves as the primary key. This table contains the details of the patients. The third table consists of the mode of admission of the patient into the hospital. Regarding relationships between the tables, an encounter belongs to one and only one patient, while a patient may be readmitted to the hospital. A few key assumptions we have made here. To begin with, we have considered only one encounter per patient, meaning if the patient was readmitted then we have not taken him/her into further consideration. Two, we have considered only those patients who were only on a single medication, Metformin. If patients were prescribed multiple drugs, then we have excluded them from the analysis. Lastly, the survivorship Bias. The patient may not choose to be admitted to the same hospital.

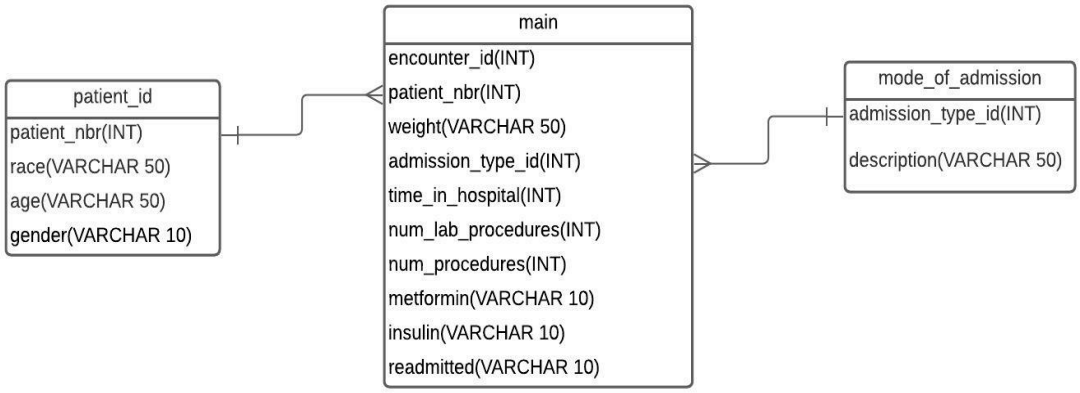

Figure 1: ER diagram

3 Extraction, Transformation and Loading

AWS Glue is a fully managed ETL(Extraction, transformation and loading) service as it makes it simple and cost-effective to store the data,clean it, enrich it and move it reliably between various data sources. Advantages of AWS Glue include:

- To build a data warehouse to organize, cleanse,validate and format data
- To run serverless queries against Amazon s3 data lake
- To create event driven ETL properties
- To understand your data source

Steps followed:

1. Collected data and divided the big data into three tables(patient id,main and ids mapping(mode of admission))
2. loaded data files into AWS S3 bucket
3. Figured out relationships among data tables via Entity-Relationship diagram
4. SQL codes were written to query the data for counting the number of males and females readmitted for a particular metformin drug, sorting the mode of admission of patients(description), sorting the age of patients using joins via Athena
5. Downloaded the ETL results and used as a input file of tableau for data visualization to display useful information in data source

4 Analysis

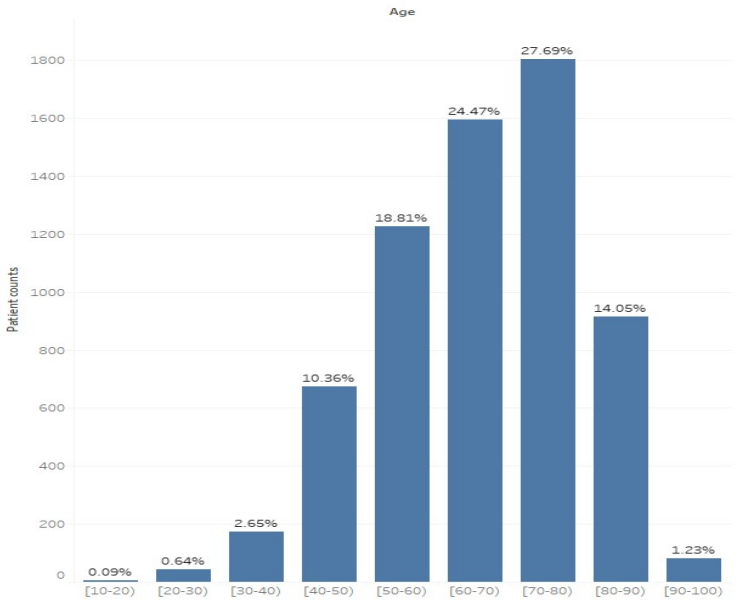

Figure 2: Age distribution

Race	[10-20)	[20-30)	[30-40)	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
AfricanAmerican	0.37%	1.78%	5.34%	18.18%	24.84%	21.56%	18.09%	9.09%	0.75%
Asian				15.00%	10.00%	15.00%	35.00%	20.00%	5.00%
Caucasian	0.04%	0.35%	1.97%	8.59%	17.56%	25.03%	29.78%	15.41%	1.27%
Hispanic		1.46%	7.30%	16.79%	21.90%	23.36%	21.17%	7.30%	0.73%

Figure 3: Classification based on age and ethnicity

Looking at the distribution of different age groups, it resembles a hypergeometric distribution. As can be seen from the graph, the readmission rates gradually increase as a person ages, which is natural given their immune system. However, there is a sudden dip in the age group 80-90. A possible explanation for this anomaly can be the fact that life expectancy in the USA is 78 years. Now, if you look among different ethnicities, this trend can be seen in general except for some oddities. African Americans catch diabetes at an early age. While

Caucasians follow the trend. Hispanics, on the other hand, have their peak at the 60-70 age group. In the case of Asians, people did not have to be readmitted until age 40. They also have the highest readmission rate of 35 percent in the 70-80 age group.

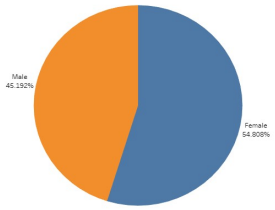

Figure 4: Genderwise classification

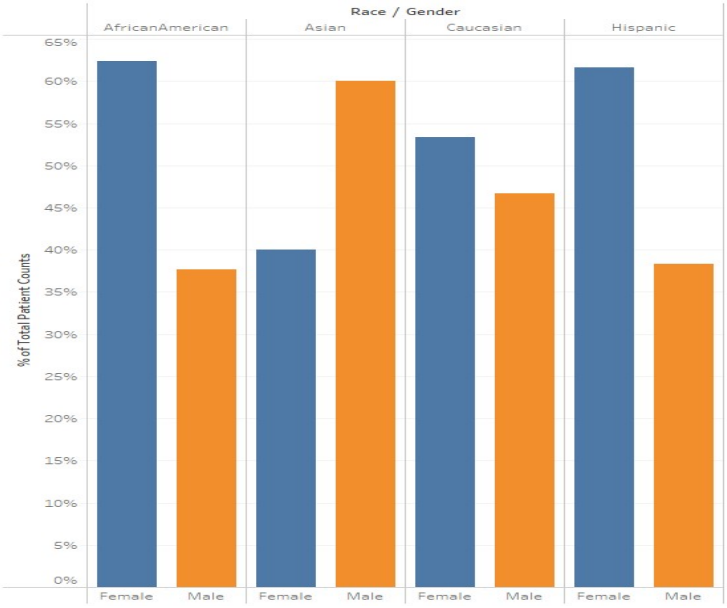

Figure 5: Classification based on gender and ethnicity

Looking at the data gender-wise, the percentage of females was 54 percent and males were 45 percent. This shows that females are at a higher risk to be affected by Metformin. Digging down the data among ethnicities, multiple things arise. The Asian race remains anomalous among the other race where the males are more susceptible to readmission than females. African Americans and Hispanics also displayed peculiarity-albeit of a different sort. The females belonging to these groups are more vulnerable post medication compared to the

males. The difference between the two sexes was a significant 20 percentage points. The Caucasians stuck to the norm.

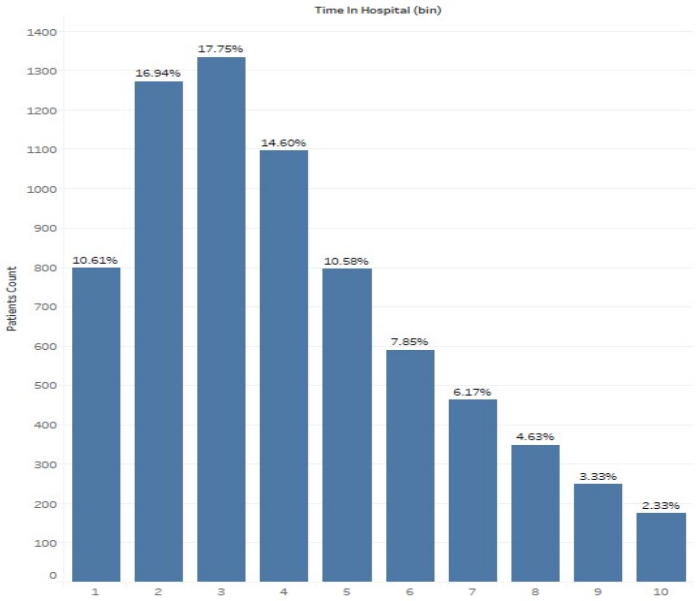

Figure 6: Average time spent in hospital

Lastly, if you look at the time stayed in the hospital after admission, it averaged to three, meaning that the patients recovered relatively quickly.

5 Conclusion

The analysis uncovers that most patients recover quickly within a week. Since most patients taking Metformin are not expected to stay over two weeks, it is safe to assume that methods used to treat the patients are effective. The most affected group is the 70-80 age group due to the natural degradation of the human body as the immune system gets weaker. It is imperative to expect more Caucasian and Asians in that age group and treat them immediately. There should be more cautious and preventative measures addressed for females since they are more susceptible for each race, excluding Asians.

References

[1] Kai McKeever Bullard, Catherine C Cowie, Sarah E Lessem, Sharon H Saydah, Andy Menke, Linda S Geiss, Trevor J Orchard, Deborah B Rolka, and

Giuseppina Imperatore. Prevalence of diagnosed diabetes in adults by diabetes type—united states, 2016. *Morbidity and Mortality Weekly Report*, 67(12):359, 2018.