Project Team 4
Members:
Chang Zhou
Harsha Jakkappanavar
Savan Patel
Jenny Johns

22 November 2017

Update 2

Project Title:
Study of dataset on Diabetes from 130 US hospitals over 10 years

Introduction/Background

Blood glucose is obtained from food which is consumed and is used as the body's primary source of energy; it is transported through the blood and brought to cells for their energy consumption ("Blood Sugar", 2017). When the level of blood glucose is elevated over time, it becomes diagnosed as the disease known as diabetes. This disease can be diagnosed with a test known as HbA1C, which looks at the average level of glucose in a patient's bloodstream over a period of three months ("Blood Sugar", 2017).

The elevation of this blood glucose can be caused by the body not making enough insulin or is not using the insulin correctly, which is a hormone from that pancreas that helps the cells of the body absorb the glucose that's travelling in the blood ("What is Diabetes?", 2016). Without it, the glucose stays in the bloodstream which is why there are then elevated levels of glucose ("What is insulin?", 2017).

The three most common types of diabetes are Type 1, Type 2 and Gestational Diabetes. Type 1 diabetes is most commonly found in kids and young adults ("Type 1 Diabetes", 2017). It occurs when the body kills its own insulin producing cells in the pancreas, thinking them to be foreign objects, also known as an autoimmune disease ("What is Type 1 Diabetes?", 2016). This means that those with Type 1 diabetes have to take insulin every day in order to manage the disease ("What is Diabetes?", 2016). Type 2 diabetes is the most common type of diabetes and occurs when the body develops what is known as insulin resistance, where the body does not properly use the insulin it generates ("Diabetes Type 2", 2016). Some patients can stabilize their blood glucose levels by eating a balanced diet and exercising regularly, but others have to take either insulin, medications, or both ("Medication", 2017).

There is no known root cause as to what causes gestational diabetes, which usually develops later on in the pregnancy, but one on-going theory is that placenta produces hormones to help the baby grow, but these same hormones may block the way insulin acts in the mother's body, causing insulin resistance ("What is Gestational Diabetes?", 2017). This not only affects the mother, but can also affect the development of the unborn baby. The elevated levels of glucose can cross the placenta and cause the baby's body to produce extra insulin in order to compensate for all the extra glucose. The glucose is then stored as fat because the developing baby does not need all that extra energy. This can lead to babies being born with the extra weight, known as macrosomia, and can mean that “the babies are born with low levels of glucose in their blood as well as being at a higher risk of breathing problems” (“What is Gestational Diabetes?”, 2017).

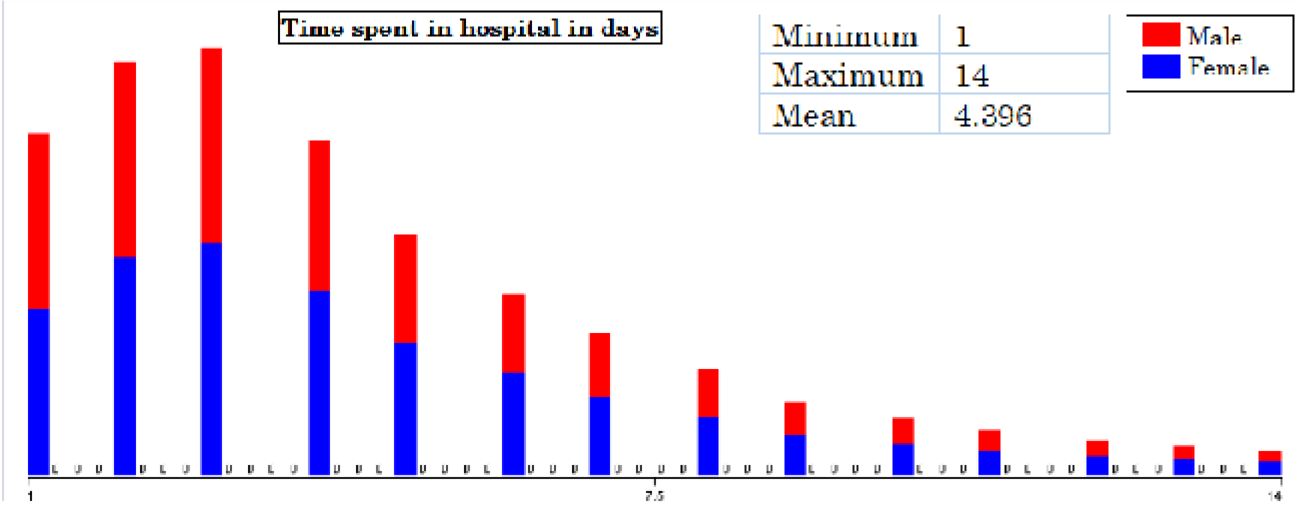
The dataset used in this project, was used in a research article, written by Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al. about how HbA1c levels affect the hospital readmission rates of patients diagnosed with diabetes. HbA1c is the measurement of the level of hemoglobin in the blood stream that has been attached to glucose. This is referred to as glycated hemoglobin (“Guide to HbA1c”, 2017). Their results included that HbA1c measurements mostly depends on the primary diagnosis. “There was a significant difference in readmission between patients whose HbA1c was checked when the primary diagnosis was diabetes and those whose primary diagnosis was a disease with their circulatory system” (Strack, DeShazo, Gennings, et al., 2014). The authors of the paper also noticed that there seemed to be a correlation between readmission and whether the Hb1Ac test was performed rather than the value of the test (Strack, DeShazo, Gennings, et al., 2014). The authors also believed that measuring the A1c would also show the effectiveness of the current treatment of care that the patient was receiving. For example, if the test resulted in >8% on the patient's current treatment, the doctors would know to adjust the medication or other therapy that the patient is on in order to account for this (Strack, DeShazo, Gennings, et al., 2014).

Exploratory Analysis

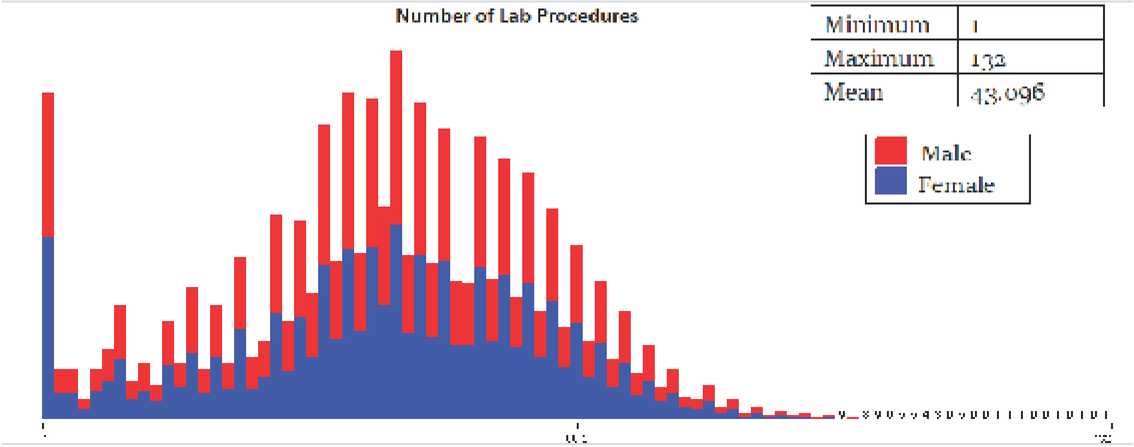
During the exploratory analysis of this dataset, we found that patients were categorized in one of five possible races, including, African American, Asian, Caucasian, Hispanic, or Other. A majority of the dataset is seen to be of Caucasian race with 76,099 out of 101,767 patient records. The second largest group were African Americans, third Hispanics, fourth Others, and finally Asians. There were some missing values from the dataset.

Gender was also broken up into three categories: Male, Female, and Unknown/Invalid. The majority of patients were female with 54,708 females versus 47,055 males. Of these patient records, there was an age range of 0 - 100, broken down into 10 year intervals. The largest subset of these patients came from the range of [70 - 80) with 26,068 patients. The distribution of patients’ weights was hard to perceive because that attribute was missing 97% of the data. The data that was added to the dataset showed that the majority of patients that whose weight was recorded landed in a range of [75 - 100) pounds.

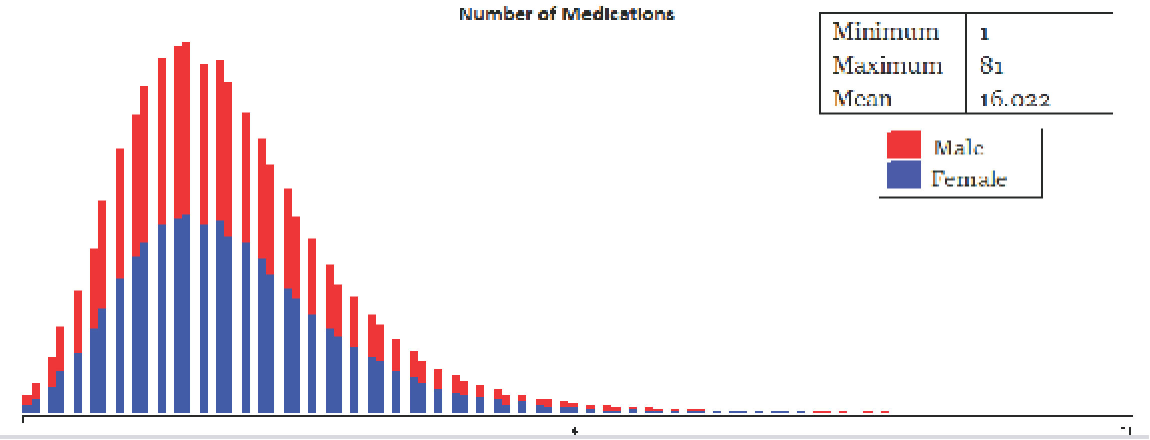
All patients were brought into the hospital with a coded admission type, ranging from 1 to 8, each representing a unique admission type. The most common admission type, was code 1, which represented an emergency admission to the facility. The time patients spent in the hospital ranged from a minimum of 1 day to a maximum of 14 days. The average time spent in the hospital was around 4 days.



Patients also had a coded payer_code which corresponded to their insurance or if the visit was self-pay. There are a total of 23 different codes possible, including insurances such as Blue Cross/Blue Shield and Medicare. While this attribute was missing 40% of it's data, a majority of the patients that did have this data coded for Medicare, with a count of 32,439 patients, which makes sense because a majority of the patients were over 65. All patients had at least one lab procedure completed during their duration at the hospital. The maximum number of lab procedures was 132 and the average amount was around 43.



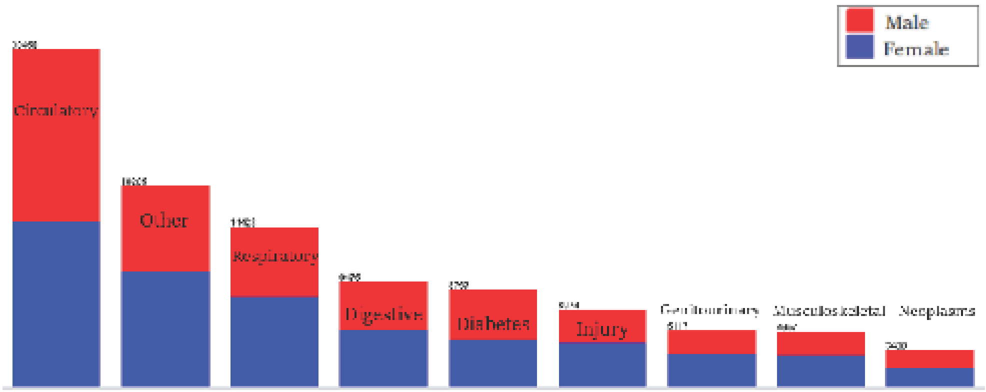
Some patients also had procedures, other than lab procedures, performed. The minimum number of procedures performed was 0 and the maximum was 6 with an average number of procedures being around 1.34. Also during their hospital stay, medications were administered to the patient, ranging from a minimum of one medication to a maximum of 81 medications, the average being around 16 medications per encounter.

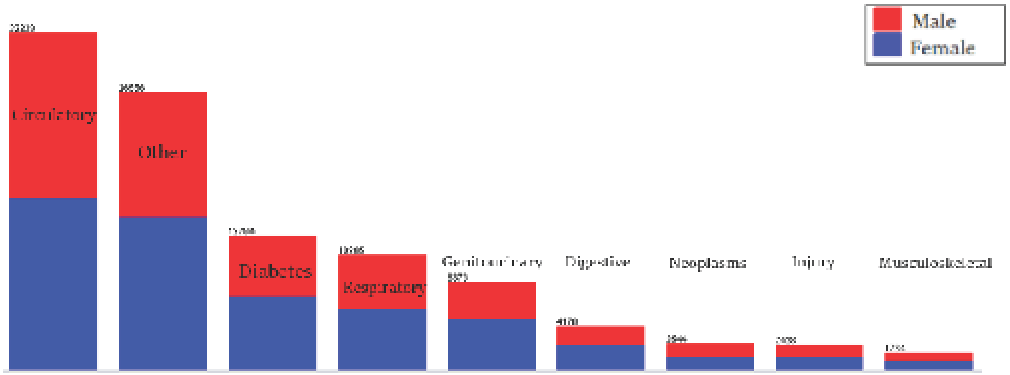


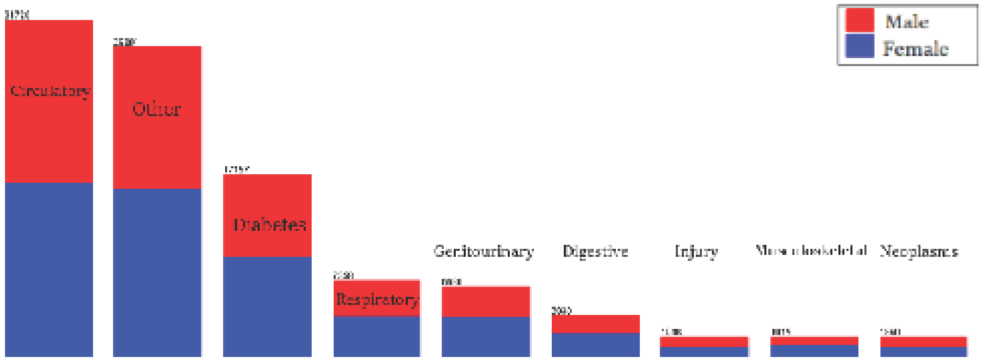
The number of emergency visits to the hospital, before this specific admission, for this patient was recorded and went back as far as a year. Not all patients had emergency visits, so the minimum was 0 while the maximum number of emergency visits was 76 within the year. The average number of emergency visits per patient was 0.198. Along with the emergency visits, it was also recorded if the patient had any inpatient visits within the past year. The minimum number of inpatient visits was 0 while the maximum was 21, with the average number of inpatient visits being 0.636.

This dataset also included the specialty of the doctor who was treating that patient. There are 84 different possibilities, including pediatrics, internal medicine, and pulmonology. This attribute was also missing 49% of data, but the patients that did have their admitting physician's medical specialty coded saw that the majority of doctors were specialized in internal medicine.

There were three different columns for diagnoses within this dataset. The first, 'diag_1' being the primary diagnosis, the second, 'diag_2' being the secondary diagnosis, and the third, 'diag_3' being an additional secondary diagnosis. Each diagnosis is a code from the icd9, representing a specific illness. Some things we noticed while working with the dataset was that a majority of the primary diagnoses were from 'Heart Failure' (icd9 code: 428) affecting 6,862 patients, from the second diagnosis, the majority of diagnoses were 'Disorders of fluid, electrolyte, and acid-base balance' (icd9 code: 276) affecting 6,752, and from the third diagnosis, the majority of diagnoses were 'Diabetes mellitus' (icd9 code: 250), affecting 11,555 patients. These diagnoses were a nominal value, and had many levels (each different icd9 code), so we grouped the codes based on which system of the body they affected, code 428, would be grouped as circulatory. Circulatory seemed to be the most highly affected system for all three diagnoses, followed by Other. The different categories were: Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Other.







The hospital could have had performed a glucose serum test on the patient. A glucose serum test shows the level of glucose in the blood of the patient at that time. The different values that this attribute can contain are ‘None’, ‘>300’, ‘Norm’, and ‘>200’. A majority of the values are ‘None’, which indicates that the test wasn’t performed. The ‘None’ value has a count of 96,420 patients, followed by 2,597 patients being tested and returning a test result of normal. We see a similar result in the A1c results of patients where a majority of patients (84,748) did not have the test performed. The different values of the A1c test can be ‘None’, ‘>7’, ‘>8’, and ‘Norm’. After the ‘None’ value the highest count was 8,216 patients who were tested and had a result of ‘>8’, which indicates that the test result shows that the A1c was greater than 8%.

There are 24 different features for medications that can be administered during a patient's hospital stay. There are four possible values for the medications within the dataset, ‘up’, ‘down’, ‘steady’, or ‘no’. For all the medications, the value with the highest percentage were ‘no’ meaning that the medication was never prescribed. Our dataset also included whether there was a change in the medications administered to the patient. The possible values were ‘ch’ for change for ‘no’ for if there was no change. A majority of the patients did not have a change in their medications, with a count of 54,755.

There was a column within the dataset that specified whether any diabetic medication was prescribed during that patient’s stay in the hospital, possible values being ‘no’ and ‘yes’. With a count of 78,363 for ‘yes’, meaning that bulk of patients were prescribed some sort of diabetic medication during their encounter. Readmission was another feature for which a record was kept. Possible values were ‘<30’ for if the patient was readmitted less than 30 days after they were discharged, ‘>30’ for if the patient was readmitted more than 30 days after they were discharged, and ‘no’ for if there was no record of readmission for that patient.

Preliminary Data Mining Analysis

PCA

This data set contains both nominal and numerical data, therefore for the numerical data, we need to normalize it and for the nominal data, we perform one-hot encoding. Before trying to run PCA, we dropped the following columns: all the 24 medication columns, 'encounter_id', 'patient_nbr', 'weight', 'payer_code', 'medical_specialty', 'discharge_disposition_id', 'admission_source_id', 'num_lab_procedures', 'number_outpatient', 'number_emergency', 'number_inpatient', and 'change'. We also change the diagnosis columns so that they represented which part of the body they affected, rather than their icd9 code. We then imported sklearn’s StandardScaler to normalize the numerical values of the data, the numerical columns consisted of: 'time_in_hospital', 'num_procedures', 'num_medications', and 'number_diagnoses'. We then applied one-hot encoding to the categorical data, which we have a manual python code for as well as using sklearn’s library for one-hot encoding, and added our normalized numerical values to create a new dataset. We passed this dataset to sklearn’s PCA and after the preprocessing of the data we were only about to get a variance of about 20%. This could be due to the fact that the majority of our dataset is categorical and PCA mainly works with numerical data.

Citations

1. Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. doi:10.1155/2014/781670
2. Blood Sugar | Blood Glucose | Diabetes | MedlinePlus. (2017, October 26). Retrieved November 04, 2017, from https://medlineplus.gov/bloodsugar.html
3. Gestational diabetes. (2017, April 28). Retrieved November 04, 2017, from https://www.mayoclinic.org/diseases-conditions/gestational-diabetes/symptoms-causes/syc-20355339
4. Guide to HbA1c. (2017). Retrieved November 04, 2017, from http://www.diabetes.co.uk/what-is-hba1c.html
5. Insulin. (2017). Retrieved November 04, 2017, from http://www.hormone.org/hormones-and-health/hormones/insulin
6. Medication. (2017). Retrieved November 04, 2017, from http://www.diabetes.org/living-with-diabetes/treatment-and-care/medication/
7. Type 1 Diabetes. (2017). Retrieved November 04, 2017, from http://www.diabetes.org/diabetes-basics/type-1/
8. Type 2 Diabetes | Adult-Onset Diabetes | MedlinePlus. (2017, November 1). Retrieved November 04, 2017, from https://medlineplus.gov/diabetestype2.html
9. What is Diabetes? (2016, November 01). Retrieved November 04, 2017, from https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes
10. What is Gestational Diabetes? (2016, November 21). Retrieved November 04, 2017, from http://www.diabetes.org/diabetes-basics/gestational/what-is-gestational-diabetes.html
11. What is Type 1 Diabetes? (2016). Retrieved November 04, 2017, from https://www.diabetesresearch.org/what-is-type-one-diabetes

Dataset

http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#