[[1]](#footnote-1)

Predicting Hospital Readmission of Diabetes Patients

Minh Quan Do

**Introduction:**

According to the Health Care Cost Institute (HCCI), total spending on healthcare per person is now growing at faster rates than prior years with a 4.6% growth in 2016 compared to a 4.1% growth in 2015 which is also an increase from just over 3% growth from 2012-2014 [1]. Furthermore, according to the National Diabetes Statistics Report, 30.3 million people in the United States (9.4% of the U.S. population) has diabetes [2]. In 2014, 7.2 million patients discharged from hospitals were diagnosed with diabetes [2]. Also, diabetes is the seventh leading cause of death in the United States in 2015 [2]. Helping hospitals know which patients are most at-risk for readmission into the hospital can help hospitals better prepare and save those more patients’ lives.

Because of data repositories such as the UCI Machine Learning Repository, statistical models have been used to analyze the large amounts of data in these repositories [3]. These techniques involve splitting the dataset up into three smaller datasets based on the patients’ age (0-29 years old, 30-69 years old, 70-100 years old) and then training different machine learning models on each of those smaller datasets [4]. The machine learning models used are a combination of decision trees, random forests, and support vector machines [4]. This technique is fairly complicated and does not take into account the heterogeneity of the patient population (at least not enough because the data is only categorized into three categories and the categories are only based on the patients’ age); therefore, the proposed method involves using clustering to organize the dataset into more homogenous groups and then using the data in those groups to train separate artificial neural networks [5]. The hypothesis is that breaking the diverse heterogenous group of patients into more homogenous groups will result in more accurate predictions and a lower error because the patients in these homogenous groups would have more similar features and theoretically, the patients will have more similar lifestyle patterns which should lead to a higher percent accuracy (Formula I) [5].

Formula I:

**Methods:**

To do this, the *Diabetes 130-US hospitals for years 1999-2008 Data Set* in the UCI Machine Learning Repository to build a dataset that would include race (nominal), gender (nominal), age (ordinal), the number of emergency visits in the year preceding the encounter (numeric), the number of inpatient visits in the year preceding the encounter (numeric), and if diabetic medication was prescribed (nominal).

McCoy et al. [6] suggests that race, gender, and age often play a factor in a person’s risk for being diagnosed with diabetes. Among the patient population, African-Americans, Asians, Hispanics, females, people who were less than 75 years old were more at risk for hospital readmission due to dysglycemia (either hypoglycemia or hyperglycemia) [7]. Due to the significance of the patients’ race, gender, and age; these four attributes will be incorporated into the model.

In addition to the patients’ physical characteristics, the patients’ medical history must be taken into consideration as well. While the overall 30-day readmission rate of hospitalized patients is 8.5–13.5 %, the 30-day readmission rate of diabetic patients is 14.4–22.7 %. Estimates of readmission rates beyond 30 days after hospital discharge are even higher, with over 26 % of diabetic patients being readmitted within 3 months and 30 % within 1 year [9]. Because diabetes patients are more likely to be readmitted, the number of emergency visits in the year preceding the encounter (numeric) and the number of inpatient visits in the year preceding the encounter (numeric) could help increase the accuracy of the model.

By far, the most common reason why diabetes patients are readmitted into the hospital is due to dysglycemia [6]. Since many diabetes patients, especially patients diagnosed with type 2 diabetes, are prescribed diabetes medication and insulin therapy to help maintain targeted blood sugar levels [10], the patient’s medical history could have a significant impact on the accuracy of the model; therefore, the model will take into consideration whether or not diabetic medication was prescribed (nominal).

Because the dataset has some missing values, data entries with missing values will be discarded to “clean” the data. Once the data is cleaned, the dataset will be split up into three portions: one portion will be used as a training set to train the machine learning model, one portion will be used as a validation, and the last portion of the dataset will be used to as the test set. The accuracy of the model will be determined using the equation described in Formula I.

Four different methods will be used for classification:

1. A model using clustering and ANN
2. Hierarchical clustering
3. K-Nearest Neighbor (KNN)
4. Changing the class labels and then applying KNN

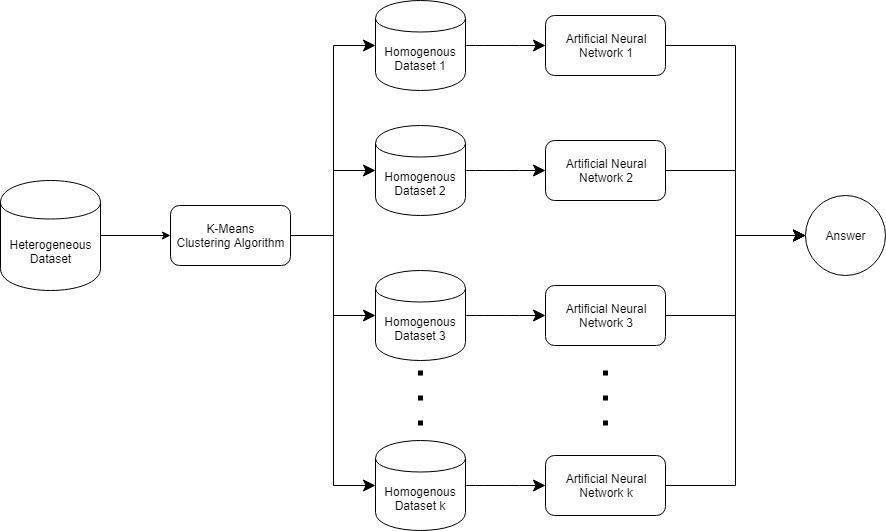
Diagram I: Flow chart of the system

*Approach I*: K-means clustering & ANN

A clustering algorithm (K-Means) will be used to organize the dataset into four homogenous groups. The rationale behind separating the dataset into these homogenous groups is that the patients’ chances of readmission is affected by their lifestyle choices [8]. Since the dataset does not include any information about the patients’ daily activities, exercising habits, eating habits, etc. clustering will be used to group patients into these groups.

Additionally, according to Hu, Gonsahn and Nerenz [11], patients who lived in high poverty (median household income of about $38,000) and low education (less than a high school diploma) neighborhoods were 28% more likely to be readmitted than those who lived elsewhere. Hu, Gonsahn and Nerenz [11] also discussed that the patients’ lifestyle choices and living conditions may be reflected in their income and education level suggesting that patients’ who live in low-income, low-educated areas would have less access to public transportation and resources such as grocery stores and pharmacies. Therefore, it makes sense to partition the dataset into four portions because there are four socioeconomic classes (upper class, middle class, working class, lower class) in the United States [12].

Once the homogenous groups have been determined, artificial neural networks will be used to analyze each homogenous group separately. The artificial neural networks would take in race, gender, age, the number of emergency visits in the year preceding the encounter, the number of inpatient visits in the year preceding the encounter, and if diabetic medication was prescribed as its inputs (6 inputs) in its input layer; and it will have three output in its output layer, which is if the patient will be readmitted into the

hospital within 6 months, after 6 months, or not readmitted at all. 

As described in Diagram I, the heterogenous dataset is sorted into K number of smaller homogenous datasets; for each homogenous dataset, a separate artificial neural network will be trained with the data in that homogenized dataset. Therefore, each artificial neural network will output a different answer. When classifying data from the test set, a point will first be categorized into a homogenous group, and then the class label will be determined using the homogenous group’s respective artificial neural network.

*Approach II*: Hierarchical clustering

Use hierarchical clustering to find group data points and then use either an ANN or a KNN to classify the points.

*Approach III*: KNN

Use training set to train KNN, use validation to find optimum K to be used by KNN algorithm, then use KNN classifier to classify data points in test set.

*Approach IV*: Change class labels and use KNN for classification

This approach involved changing all data points with class labels of “>30 Days Readmitted” to “Not Readmitted”. From here, the goal of the project is to classify if patients are to be readmitted within 30 days.

**Results:**

The original dataset had 101,776 data points. After cleaning the data, the dataset was reduced to 99,493 data points. The dataset is then divided into portions, a training set consisting of 34,000 data points, a validation set consisting of 34,000 data points, and a testing set consisting of 33,764 data points.

After testing the original approach (detailed in Diagram I and Approach I) did not perform as well as expected so three other approaches were also used to classify the dataset.

*Approach I*: K-means clustering & ANN

To train the model (Diagram 1), the K-means clustering algorithm was employed (K = 4) to partition the training set into four clusters. The clusters Group1, Group2, Group3, and Group4, had 2,071 data points, 20,155 data points, 5,577 data points, and 6,197 data points respectively.

Once the training set was partitioned into four clusters, four artificial neural networks ANN1, ANN2, ANN3, ANN4 were trained using data from Group1, Group2, Group3, and Group4 respectively. The artificial neural network has no hidden layers, only an input layer with 6 inputs and an output layer with 3 outputs.

The resulting confusion matrices displaying the accuracy of the four artificial neural networks is shown below in Figure 1-4:

Figure 1: confusion matrix representing training accuracy for ANN1

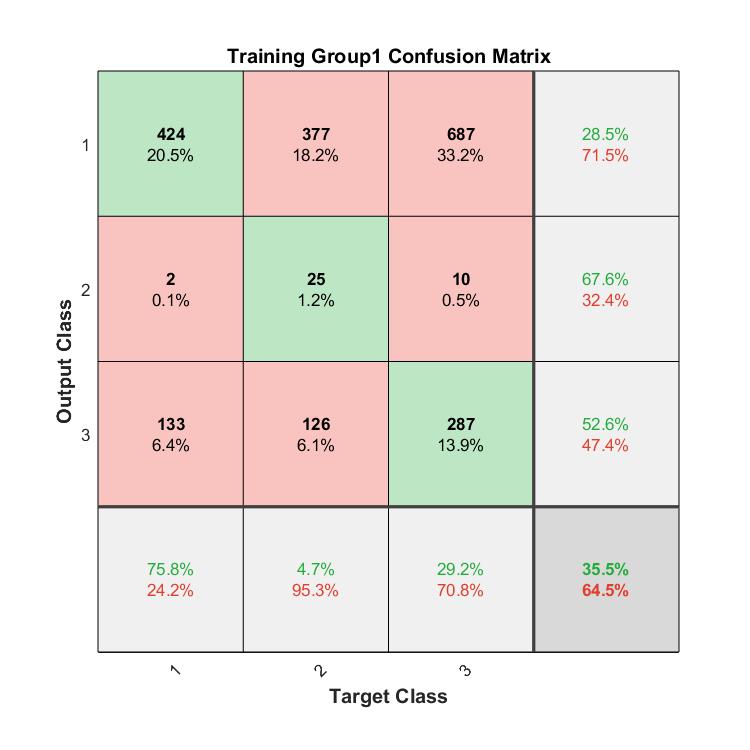


Figure 2: confusion matrix representing training accuracy for ANN2

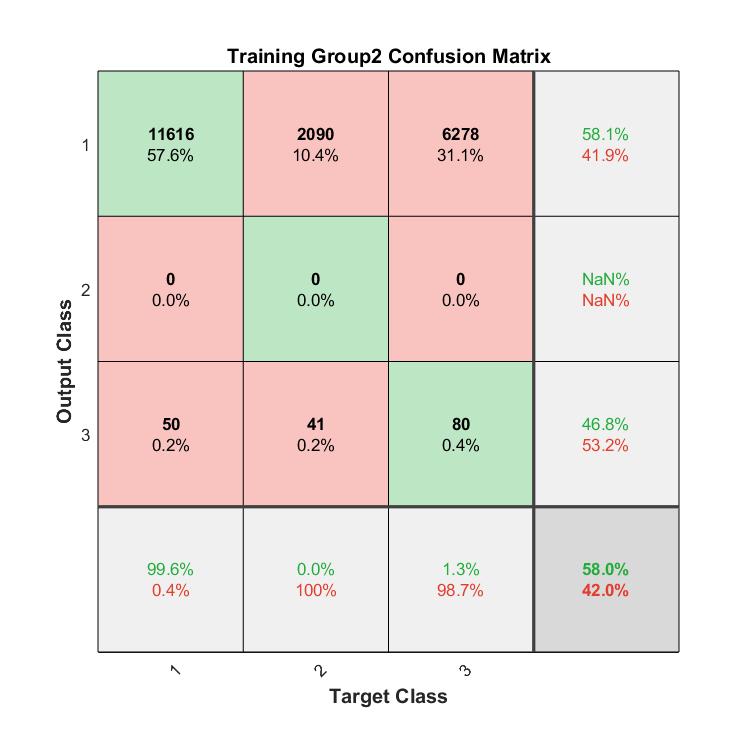


Figure 3: confusion matrix representing training accuracy for ANN3

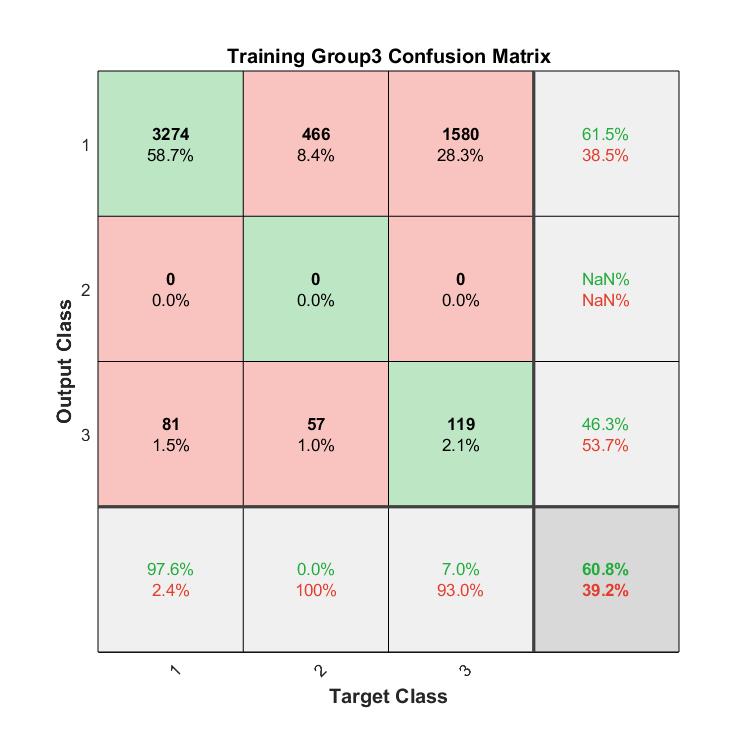
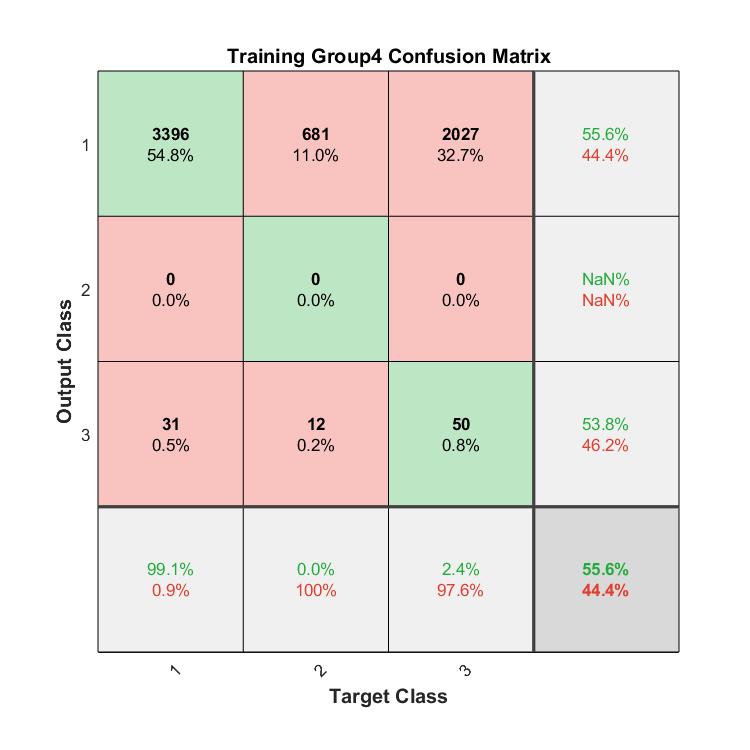


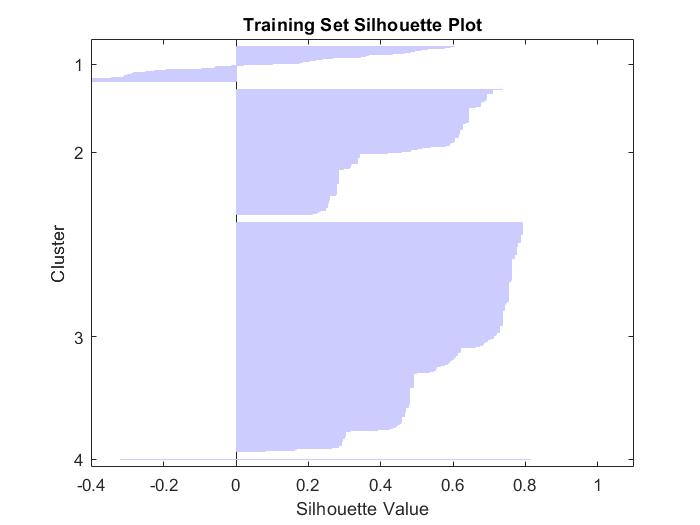
Figure 4: confusion matrix representing training accuracy for ANN4



As stated, these confusion matrices represent the training accuracy of the four neural networks after the model has been trained on the training set. According to the confusion matrix for ANN1 (Figure 1), the training accuracy is not very good as the neural network only classified 35.5% of the points correctly. ANN2, ANN3, and ANN4 correctly classified 58% of the data points on average. While the training accuracy for ANN2, ANN3, and ANN4 are still quite poor, it is indeed much better than ANN1.

It should be noted that the number of points assigned to a cluster varies greatly every time the K-means clustering algorithm is ran; therefore, the training accuracy could also vary greatly between each trial.

Figure 4: silhouette plot of K-means clustering algorithm

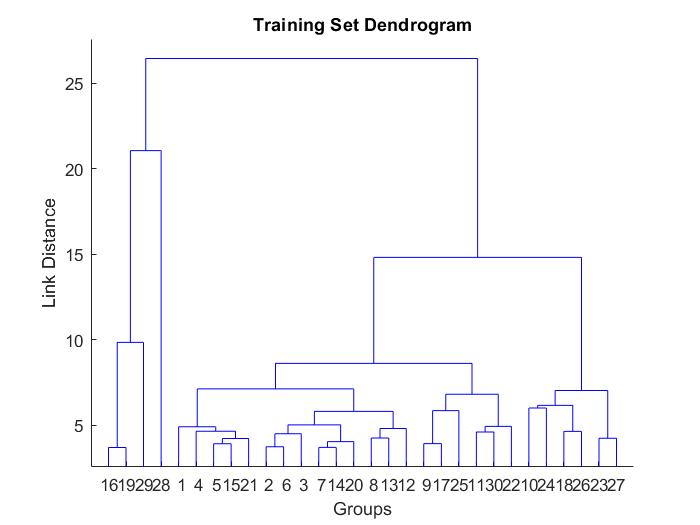


The reason why the number of points assigned to the clusters vary is because there were many misclassified data points as evidenced by the negative silhouette values in the silhouette plot in Figure 4. The most likely cause for the numerous misclassified data points is because the dataset is not separable.

*Approach II*: Hierarchical clustering

The second approach tried was hierarchical clustering. Unfortunately, hierarchical clustering yielded results that were no better than K-means. The most likely reason for the failure of hierarchical clustering is probably because there were many misclassified data points which is cause by the dataset being misclassified which is also the same reason that resulted in the failure of K-means.

Figure 5: Dendrogram



When looking at Figure 5, the dendrogram shows that a linking distance of around 13 is a good cutoff point and it seems that the dataset can be broken up into roughly four clusters. However, because the data is inseparable, hierarchical clustering is not an effective means to separate the data points and increase the accuracy of the model. Therefore, since hierarchical clustering did not work, there was no reason to continue with attempting to classify the data points in the test set using this approach.

*Approach III*: KNN

Another approach that was tried was using KNN. KNN was used to classify the points in the validation set and the test set. The validation accuracy and the testing accuracy were about 50% which is not very good. Additionally, finding the optimal number of nearest neighbors also proved to be an very computationally expensive task; therefore, KNN is not a suitable option to classify this data set.

*Approach IV*: Changing the class labels & KNN

When the class labels were changed, the training set now consists of 30,125 data points labeled “Not Readmitted” and 3,875 data points that are labeled “<30 Days Readmitted”.

When using KNN to classify the data points, setting K = 1 results in a testing accuracy of 80.5%, a sensitivity of 17.11%, and a specificity of 88.23%. If K > 15, the accuracy would increase to over 89%; however, the sensitivity would decrease to less than 1%. Therefore, it is determined that it is best to set K = 1.

The most likely reason for the loss of sensitivity is due to the model misclassifying “<30 Days Readmitted” data points as “Not Readmitted” because there is an overrepresentation of “Not Readmitted” data points in the training set. Overrepresentation of “Not Readmitted” data point is probably also what lead to the large increase in testing accuracy because less “Not Readmitted” data points are being misclassified as “<30 Days Readmitted” or “>30 Days Readmitted”. Overrepresentation with inseparable data also explains why the sensitivity of the data decreases as K is increased. As K increases, the probability of the classifier detecting more data points labeled “Not Readmitted” increases; and since there are not nearly as many data points labeled “<30 Days Readmitted” to begin with, it is much more likely that the classifier will categorize points as “Not Readmitted”.

**Conclusion:**

In retrospect, the dataset was determined to be very convoluted and inseparable and therefore very hard to perform classification. Although the changing the class labels can dramatically increase validation/testing accuracy and specificity, it comes as a cost of decreased sensitivity. This decreased sensitivity is a concern because it means that the model is not very reliable in predicting the risk of readmission of patients.

Diagram 2: Damian Mingle’s model

In the future, before attempting to perform classification on this dataset, principle component analysis (PCA) should be used to determine the right parameters to include in the model. Additionally, finding the p-values of each parameter could also help to determine the most significant parameters to be considered in the model. A possible way to increase the sensitivity of the model is to use an ensemble learning method that incorporates the predictions of many different classifiers to improve predictive performance like Damian Mingle’s model that was also used to predict readmission rates of diabetic patients (Note: Mingle also used the same data set and the model represented in Diagram 2 obtained a testing accuracy 84.81% and a sensitivity of 49.78%) [4].

Unlike Damian Mingle’s model which involved splitting the data set up according to the patients’ age, it might be better to use hierarchical clustering to cluster the data points according to the information presented in the data set to hopefully get more meaningful, more homogenous groups of data points, which would hopefully lead to a higher testing accuracy and sensitivity. Additionally, using PCA and p-values could also help in adding parameters to the model which could help to separate the data due to the increased dimensionality.

Although this model may not have been the most successful, it is possible that as time goes on, more data will be added to the dataset due to better record keeping practices and more patients being readmitted; and as a result, the extra information gathered can help future researchers to build better more effective models based on the ideas and lessons presented in this project.

References

1. United States, Congress, “2016 Health Care Costs and Utilization Report.” *2016 Health Care Costs and Utilization Report*, 2016.
2. United States, Congress, Division of Diabetes Translation. “National Diabetes Statistics Report.” *National Diabetes Statistics Report*, 2017.
3. Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “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.
4. Mingle, Damian. “Predicting Diabetic Readmission Rates: Moving Beyond HbA1c”. *Current Trends in Biomedical Engineering & Biosciences*, 2017, 7. 10.19080/CTBEB.2017.07.555715.
5. Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J Garcia-Laencina, Adelia Simao, Armando Carvalho, “A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients”, *Journal of biomedical informatics*, 58, 49-59, 2015.
6. McCoy, R., Lipska, K., Herrin, J., Jeffery, M., Krumholz, H. and Shah, N. (2017). Hospital Readmissions among Commercially Insured and Medicare Advantage Beneficiaries with Diabetes and the Impact of Severe Hypoglycemic and Hyperglycemic Events. *Journal of General Internal Medicine*, [online] 32(10), pp.1097-1105. Available at: https://link.springer.com/article/10.1007%2Fs11606-017-4095-x [Accessed 19 Nov. 2018].
7. Kautzky-Willer, A., Harreiter, J., & Pacini, G. (2016). Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus. *Endocrine reviews*, *37*(3), 278-316.
8. Mayo Clinic. (2018). *Diabetes - Symptoms and causes*. [online] Available at: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444 [Accessed 18 Nov. 2018].
9. Rubin, D. (2018). Correction to: Hospital Readmission of Patients with Diabetes. *Current Diabetes Reports*, [online] 15(4). Available at: https://link.springer.com/article/10.1007%2Fs11892-018-0989-1 [Accessed 19 Nov. 2018].
10. Mayoclinic.org. (2018). *Type 2 diabetes - Diagnosis and treatment - Mayo Clinic*. [online] Available at: https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/diagnosis-treatment/drc-20351199 [Accessed 19 Nov. 2018].
11. Hu, J., Gonsahn, M. and Nerenz, D. (2014). Socioeconomic Status And Readmissions: Evidence From An Urban Teaching Hospital. *Health Affairs*, [online] 33(5), pp.778-785. Available at: https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2013.0816 [Accessed 19 Nov. 2018].
12. Open.lib.umn.edu. (2018). *8.3 Social Class in the United States – Sociology: Understanding and Changing the Social World*. [online] Available at: https://open.lib.umn.edu/sociology/chapter/8-3-social-class-in-the-united-states/ [Accessed 19 Nov. 2018].

1. [↑](#footnote-ref-1)