Predicting Occurrence of Chronic Diseases in the Medicare Population:

James Perry, Robert Steele, Patrick Magnusson, Jonathan Worden

ABSTRACT

Health care spending in the United States represents one of the largest single budget categories in our annual spending. The proportion and scope of medical spending has been growing at an increasing rate for several decades, largely in part due to lifestyle changes in our population, as well as advances in medical technology that drives the costs of advanced treatments and drugs up; not to mention government policy changes. Preventative care business models focus on reducing cost by heading off potential conditions before they progress into extreme and expensive cases. Deep neural networks can be implemented in such business models in order to yield fast and reliable results to help predict patients who might be at a higher risk for a disease based off currently unlinked patient condition data. These example exploratory neural networks are built off of actual medical data gathered over several years and cover a large array of possible medical conditions ranging from diabetes to nervous system ailments. The effective use of predictive models could yield better medical service to the patient, decreased cost to the business, and possibly attract more medical providers to adopt preventative care, decreasing health care spending.

Results of this exploration are mixed and show that given the right conditions, such a solution is well-suited to predicting conditions that correlate strongly with others. However, those conditions which are less so would be better served with a different predictor/input set and/orother testing methodology.

I. Introduction

In healthcare, it is becoming increasingly important to lower the cost of care overall. As of 2016, 17.9% of the United States of America’s gross domestic product (GDP) was spent on health care costs. This makes health care one of the USA’s largest industries, totaling about $3.3 trillion dollars [1]. As a comparison, in 1960, only 5% of the USA’s GDP was attributed to health care cost [1]. This trend in rising cost is not sustainable for our country. The need for preventative medicine practices is paramount to buffering the rising costs and also helps provide better quality care for our citizens. Preventative medicine is becoming more widely used, especially on the Medicare population, largely in the form of early testing for diseases. The Medicare population of members is especially interesting to the problem of rising health care costs simply because this population is largely made up of elderly individuals 65 years of age and older. The early disease testing done for these members is helping to reduce cost by identifying problems in patients before they become uncontrolled and hospitalizations or death occurs as a high consequence of the disease going undiagnosed.

Through modern techniques in modeling, it has become possible to try and predict the existence of certain diseases for individuals, given the data is sound enough to utilize and has some signal to be captured. We aim to prove that an applicable set of models can be built to help identify suspects for key chronic diseases in the Medicare population to aid in preventative medicine efforts in the elderly sector of medicine. This will not only lower healthcare costs, but also increase quality of life for this sector of the US population as they head into their golden years. The diseases we will try to predict are: Diabetes with complications (industry HCC code 18), Morbid Obesity (HCC 22), Neuropathy (HCC 75) and Chronic Obstructive Pulmonary Disease (HCC 111).

II. Data Structure and Modeling Approach

*1. Data Structure*

The dataset is a collection of labeled patient data ranging from 2014-2017. This data is void of any HIPAA sensitive fields: consisting of only a row ID, Year of Data Indicator, Age, Gender and binary flags for all Disease Classes (HCCs) as specified by the Center for Medicare-Medicaid Services (CMS), the governing body for Medicare. The overall data set consists of just over 818,000 rows. The idea behind the construction of this data set was to use as little information as possible about the patient, other than some basic demographic factors like age and gender to see if a signal exists for predicting the existence of a specific condition, given you have a list of all other known conditions the patient has, a concept referred to as comorbidities. CMS generally trains their predictive models on one year, and then tests the results on the following year. We will try to follow this same protocol in our modeling in the hope of matching this industry standard.

TABLE I

Disease Prevalence By Year - WellMed Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | Diabetes With Complications | Morbid Obesity | Neuropathy | Chronic Obstructive Pulmonary Disease |
| 2014 | 40% | 17% | 8% | 28% |
| 2015 | 42% | 18% | 10% | 29% |
| 2016 | 42% | 19% | 11% | 29% |
| 2017 | 43% | 20% | 13% | 29% |
| Total Prevalence | 42% | 19% | 11% | 29% |

Table 1 shows the prevalence of each disease state we are trying to predict. Prevalence is a measure of the percentage of the population that has been diagnosed with a particular condition. In our models, it is important to keep in mind these prevalence values when evaluating the accuracy of the model overall. We need to make sure that we are outperforming the proportion of the larger of the two possibilities (0 or 1) overall, as anything less this results in a model inferior to merely guessing the more likely option.

*2. Modeling Approach*

Each of the four conditions chosen are tested separately. All of the models built for the disease prediction are based on feedforward DNN neural networks. Because a patient either does or does not have a specific disease, the response becomes a binary situation from a predictive standpoint. Due to each condition having its own likely predictive signal in the data, and correlates of interest, we decided that a single model to rule them all probably wouldn’t work. As such, there are 4 separate modeling attempts, 1 for each condition we were trying to predict. The models were all trained and validated using a specific year’s data (whichever was most complete), and then the models were tested on the following year’s data for overall accuracy.

The metric we are trying to maximize to indicate a good model is positive predictive value (PPV). We believe that creating our models to maximize this value allows us to fine tune our true positive rate so that the conditions can be suggested to providers without too much provider abrasion (improper diagnosis or procedure leading to insurance claims issues). It is most important to correctly diagnose a true positive and avoid false positives. We don’t want to waste a provider’s valuable time chasing a diagnosis on a non-present condition, so we believe maximizing the PPV, rather than the overall model accuracy, will help to alleviate this issue. However, as will be seen, even in cases where the PPV is very high, there can be issues with the overall picture of accuracy remaining.

III. Modeling Results

1. *Diabetes with Complications*

Diabetes mellitus, more commonly referred to as simply diabetes, is a chronic disease that is becoming more and more prevalent in America. According to the World Health Organization, “Diabetes mellitus is a chronic disease caused by inherited and/or acquired deficiency in production of insulin by the pancreas, or by the ineffectiveness of the insulin produced. Such a deficiency results in increased concentrations of glucose in the blood, which in turn can damage many of the body's systems, in particular the blood vessels and nerves [2].” It is important to catch diseases like this early. In our case, we are only looking for diabetes with some kind of complication in terms of the disease class (there is a separate code for non-complication diabetes, which will not correlate as strongly).

As seen in Table 1 above, the prevalence rate for diabetes with complications in our data set is about 42% overall. So, this means that we need to target an overall accuracy rate of about 58% before the model becomes of any incremental use. 2016 was the year chosen for modeling as it was the most complete and had the most rows of data. The way the Medicare business works, a dataset usually isn’t complete until after about two years. The way the workflow happens, data can continue to change up until around then. The 2016 training set consisted of about 248,072 rows overall and it was split 70%/ 30% for training and validation, respectively.

The model topology for HCC 18 prediction underwent many iterations of tweaks before settling on a final model. The final model was sequential in nature and consisted of four layers total. Research to pin down which activation functions might be best for our prediction types, found that for HCC 18 prediction, the simple relu and sigmoid proved to give good results. We believe that for our problem, the relu activation function served to filter out a lot of the noise that existed within the data quickly [3]. Given the size of our dataset and the nature of the business, we need quick results to get the information out to the provider, so care can be rendered for the patients, and having a model that is efficient is key to getting results fast. Relu seems to help with that and this was evident from run times seen in the models. In the last layer of the model, a sigmoid activation function was used for two main reasons: 1) sigmoid will squeeze values between 0 and 1, which makes everything into a probability, 2) the sigmoid function is an inherently good classifier, which we need to classify our patient as either 0 (they don’t have the disease) or 1 (they do have the disease) [3]. These approaches are maintained across each of the subsequent conditions modeled (save for some variation in layer structure).

The input and layers shapes were tuned so that the total parameters trainable were roughly the same size as the training dataset itself. Since our dimension size was 83, this means starting off with around 500 for the initial squeeze, and we cut that number in half for the most part down to 1 in the final layer.

Just using these parameters here and considering nothing else, we were able to achieve about 85% accuracy on the training set. However, on the validation set, things got interesting. The training went well, but the model did not learn very well at all in a manner that transformed to a validation set. Figure 1 illustrates this.

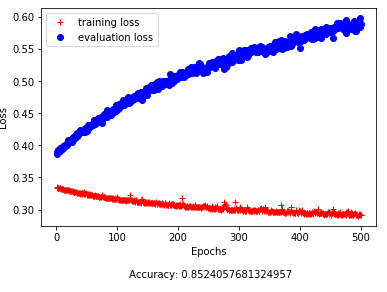


Figure 1. Initial model topologies showing inability to generalize to validation set

After this result, different methods were applied; everything from changing the activation functions all to sigmoid, to trying kernel initializers like random normal initializers. None of these seemed to improve the models overall. Eventually, we got to a model that produced Figure 2, which is a classic example of overfitting.

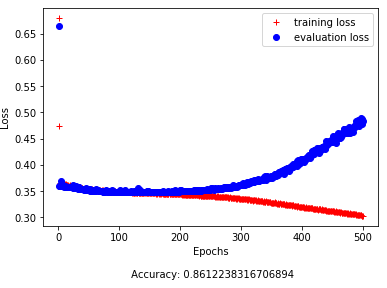


Figure 2. Intermediate model stage showing evidence of overfitting on the training set

After researching how to mitigate overfitting in our models, kernel regularlizers were applied to add a penalty in the loss function [4]. The regularizer that seemed to work best for this problem was an L2, which is a penalization method that reduces the weights in the model under the assumption that smaller weights lead to simpler models. This is supposed to help with overfitting [5]. This proved to be quite helpful, producing the final model seen in Figure 3. Note that Figure 3 was also run on 1500 epochs, with the best model being found 726 epochs in.

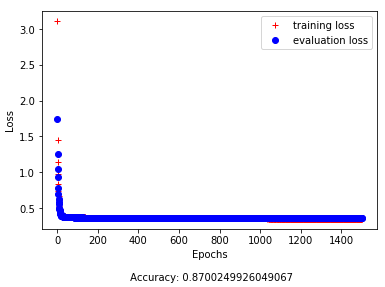


Figure 3. Final model producing the best accuracy for both training and validation

Now that we found a pretty good model, it became important to stress test it and determine whether it could generalize well. We have a model that can estimate the occurrence of diabetes with complications at about 87% accuracy, which is pretty good considering prevalence rate is again at 42% in the population. So, we are far exceeding that 58% break even point. Following with CMS protocol, we decided to apply the best model to the 2017 data set as a final test. The overall accuracy fell to about 81.33%, which was disappointing to see, but it doesn’t tell the whole tale from our perspective. Because we are focusing on PPV, not accuracy as the metric we want to maximize, we can still fine tune the results.

TABLE 2

Confusion Matrix of Final Model for HCC 18

|  |  |  |  |
| --- | --- | --- | --- |
| n = 280,158 | Predicted: 0 | Predicted: 1 |  |
| Actual: 0 | 157,149 | 2,865 | 160,014 |
| Actual: 1 | 62,255 | 57,889 | 120,144 |
|  | 219,404 | 60,754 |  |

Table 2 shows a probability cut on the predicted values greater than 90%. This gave us a PPV value of 95%, which we thought would get great buy-in from an executive that might have to consider using a model like this. The only problem with maximizing the PPV, is the possibility of introducing more false negatives by maximizing the true positives. In this case, we are leaving 22% in the false negative bucket, which may or may not be a good play from a care opportunity perspective. The PPV value in conjunction represent a dance that we will have to play in terms of how comfortable is one recommending a suspect condition to a doctor and it not be a true positive. The more false positives the doctors get, the less buy-in there will be on the model recommendations.

1. *Morbid Obesity*

Morbid Obesity is a condition in which “excess fat accumulation (regionally, globally, or both) increases risk to health [6]”.A common used metric to assess severity and degree of obesity, whether generalized, extreme, or morbid (which is a resultant condition of having one of the prior two) is the Body Mass Index, which is a function of height and weight (BMI). Generalized obesity is considered a BMI > 30 kg/m^2, and extreme obesity when BMI > 40 kg/m^2 [6].

Morbid Obesity is a rising trend in the United States, especially among minorities and senior citizens [6]. Morbid obesity is closely linked with other health conditions that develop due to being overweight, such as Type 2 Diabetes Mellitus (see prior section), and many heart conditions [6]. To provide the best care possible, weight loss can be a powerful preventative technique to reduce risk of other conditions developing. A model to aid in diagnosing this condition can be a precursor to assigning preventative care, so as to avoid future complications.

Our data shows a rising trend in this condition (denoted as HCC 22) from 2014 to 2017, going from 17% to 20% prevalence in the WellMed patient pool in only 3 years. However, it is very important to note that this is substantially below the prevalence in the United States adult population at large. Table 4 illustrates the gaps between the national prevalence rates (as denoted by the CDC), in adults 18+ and our dataset. This gap in prevalence will increase sensitivity requirements for the condition in any categorization modeling performed.

TABLE 3:

Prevalence Gap in Morbid Obesity Condition [7]

|  |  |  |
| --- | --- | --- |
| Year | CDC National Trends (18+) | WellMed Dataset |
| 2017 | N/A | 20% |
| 2016 | 29.6% | 19% |
| 2015 | 28.9% | 18% |
| 2014 | 28.9% | 17% |

A feed-forward DNN neural network was constructed on 2016 data to be tested on the 2017 data. 2016 was the most complete year from a data perspective as it was directly after a company data integrity review, and as such the data was determined to be the most trustworthy to train a model on. A 70%/30% train/test split was used on the data, with the 30% of 2016 data held out as a validation set. 2017 was to be the full test set.

The initial neural network model constructed had the model topology and results shown in Figure 4. These results were very concerning for several reasons. 1) PPV (our metric of choice as discussed in the prior section) is very poor at 56.19% and many false positives. 2) The loss tracking chart displays powerful signs of the model overfitting on the training data and not generalizing to the test data. 3) Overall validation accuracy is right on with the proportion of non-conditions (0s) in the dataset, indicating that the model is not learning to find the condition and is no better than a guess. This is verified by the fact that there are very few true positives found, certainly not the true proportion of 20%.

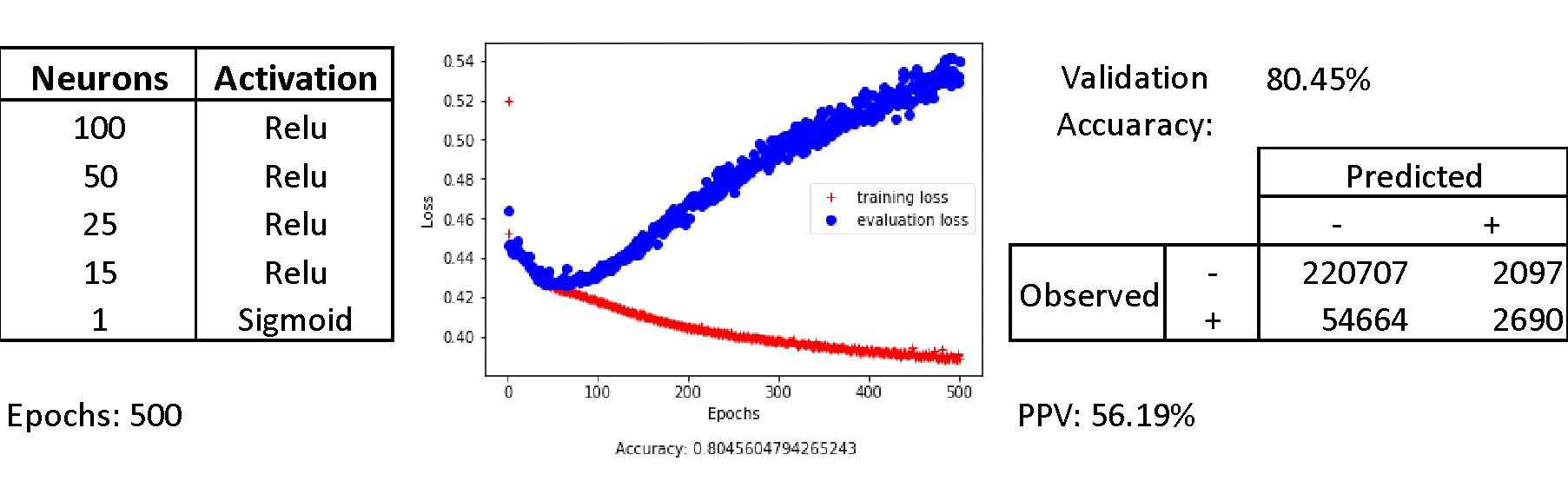


Figure 4: Initial Obesity Model

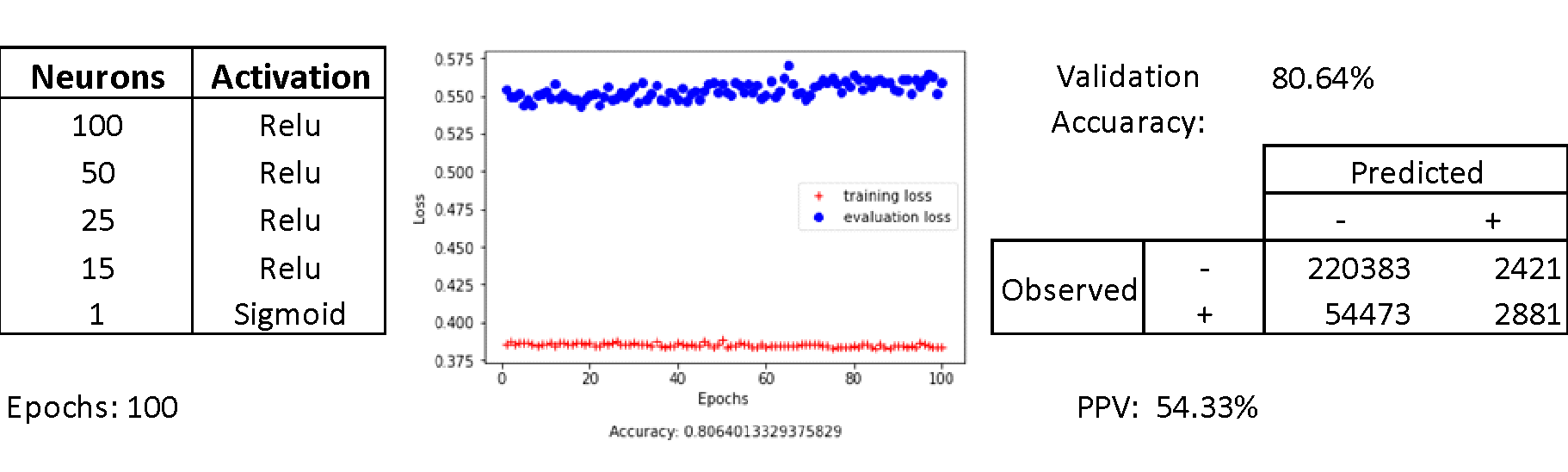
In order to combat this, a short model of only 100 epochs was constructed to zoom in on the inflection point and avoid the point of overfit. However, this was ineffective (figure 5) as the error bars do not even converge at this few epochs in the network, and the accuracy and PPV are not improved.

Figure 5: Short Model

The same set of challenges and problems plagues the model when many methods are used and iterated upon in attempt to resolve the lack of learning in the network (figure 6). The following methods were tested extensively:

1. Changes in learning rate,
2. A lower number of learned parameters (simpler network)
3. A lower-epoch version of the simpler network
4. Kernel Regularization
5. Removal of all-0 rows to reduce sparsity and promote signal on HCC 22

All of the methods in Figure 6 failed to mitigate the effects of overfit, lack of error bar convergence, and poor PPV. The conclusion reached from this is that the condition is extremely difficult or not trainable to predict solely based on the variables provided in the data. This assertion is supported by the effects of the kernel regularization, which cause the model to be unable to determine any 1’s (condition present) at all, whether true or falsely positive. The penalty on the model incurred by the L1 and/or L2 methods are stronger than any signal present in the dataset, implying that the few observations captured correctly in other passes of the model are a fluke more than any accurately trained method. The modeling efforts may also be negatively impacted by the gap in prevalence between this condition and the national true prevalence, which will make categorization more difficult for the model.

This inability to predict makes some logical sense as well, as diabetes, the first condition predicted, is correlated strongly with other conditions, while obesity is more loosely if at all correlated with many of the conditions in the dataset, save the few that it can be a precursor to causing [6]. Not only this, but given the simplicity of diagnosing this condition using a simple BMI ratio, a medical professional is unlikely to use a modeling approach such as a neural network to predict this condition. WellMed should instead seek to build models on conditions where diagnosis is more difficult. Nonetheless, this modeling provided a valuable lesson in this regard.

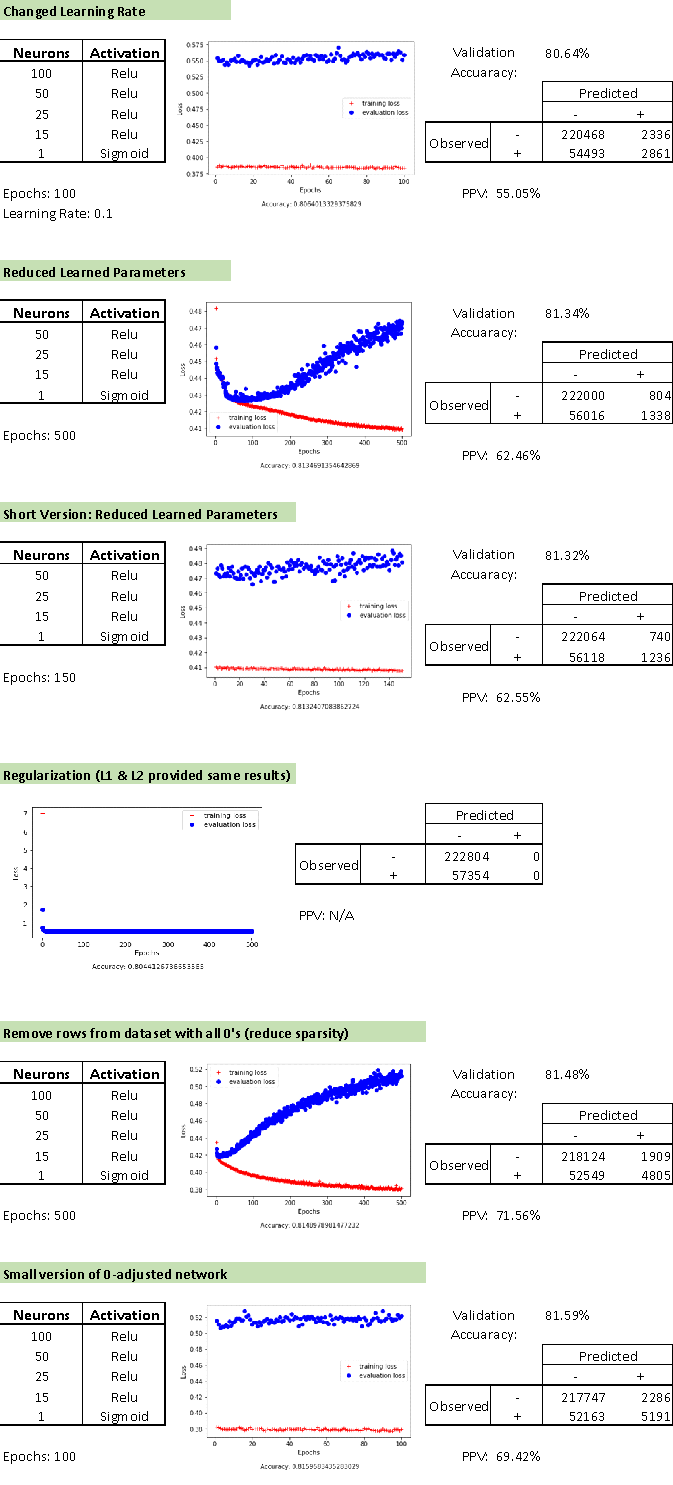


Figure 6: Various Mitigation Methods

1. *Neuropathy*

Neuropathy is a term used broadly to describe a variety health problem concerning the nervous system. The leading causes that induce said damage are diabetes (60%), idiopathic (23%), and chemotherapy (10%). There are between 20 and 30 million Americans that suffer from some form of neuropathy [10]. The symptoms can vary depending on the cause but may include; tingling or numbness in the nerve, muscle weakness, paralysis, organ dysfunction. Of the largest group (diabetic neuropathy) nearly 54,000 have amputations as a result of neuropathy; an estimated 75% of these are preventable. The costs for just this one type of neuropathy pushes 58 billion dollars a year[11].

There is no cure that will completely reverse the damage done by neuropathy. However, there are several things a person can do to prevent and/or speed up recovery such as diet, lifestyle, and treatment changes. Treatment revolves around the management of symptoms. This usually takes the form of medications such as antidepressants, antiseizure, an antiepileptic medication that have an effect on the specific neurotransmitters. It is equally important to cease the inciting activity; additional damage can drastically slow down a healing process that can take 18 months to several years to recover [11].

Neuropathy is represented in the data by ‘HCC 42’ and will serve as the predictor for the model. We will be shooting for over 88% accuracy as neuropathy has a prevalence of 11% within the dataset. Heavy consideration is given to PPV over accuracy given the negative effects and cost of false positives in diagnosis of medical conditions. In the model building process several different approaches were attempted. Early results demonstrated the greater performance of relu over sigmoid or tanh hidden layers with the exception of the final layer for reasons state previously. Models in the early stages consisted of 5 layers starting at 500 neurons and the uniform initializer. Many alternate initializers were attempted but accuracy began to lower significantly. This initial models’ best iteration at 88% accuracy with a PPV of 59% (figure 7) and 500 iterations. Unfortunately, the testing process demonstrated the glaring failure of this model to generalize to the testing set.

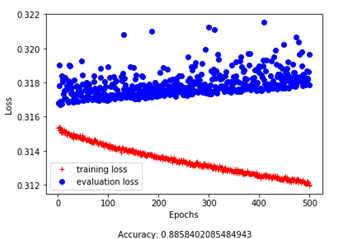


Figure 7: Initial Neuropathy Model

There was more success in the use of kernel regularizers to prevent overfit. While both L1 and L2 regularizes reduce overfitting L1 (Gaussian) is less tolerant of weights further from 0 than L2 (Laplace) [4]. All models using these regularizes arrived at 88% accuracy, however, all models including L1 regularizes failed to yield a single true positive or true negative. At this state the best model yielded 88% accuracy with a PPV of 44.87% (figure 8) and 2000 iterations. Given our prevalence and the low accuracy at this stage it would be more effective to simply assume negative presence of HCC 42. This is the same phenomenon exhibited when modeling morbid obesity.

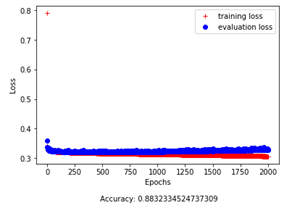


Figure 8: Neuropathy Model with Regularization applied

Subsequent changes focused on reducing the layer count and width of neurons at a given layers in order to not overwhelm the model with learnable parameters. Multiple sized runs were conducted all the way to the minimum of 82 neuron width in the initial layer. This was much more successful reaching a PPV of 67% and accuracy of 88% (fig N 3) after 500 iterations. However, false negative count remains very high, and as with morbid obesity, the model does not reach the standards required by practical usage to predict the condition beyond current medical procedures.

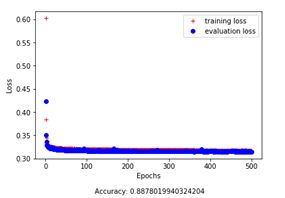


Figure 9: Low-Parameter Neuropathy Model

Given the failure of these models to overcome the minimum accuracy of 90% and to reach a reliable PPV it is unlikely that this model could be implemented to the betterment of the patient on behalf of the medical practitioner.   
.

## *Chronic Obstructive Pulmonary Disease (COPD)*

The American Lung Association (ALA) describes COPD as follows:

Chronic obstructive pulmonary disease (COPD), which includes chronic bronchitis and emphysema, is a long-term lung disease that makes it hard to breathe. The disease is increasingly common, affecting millions of Americans, and is the third leading cause of disease-related death in the U.S. The good news is COPD is often preventable and treatable. [8]

One of the challenges with this disease is that it is often not found until it is very advanced. While COPD is not curable, if found early, much can be done to treat and help the disease.The current method of diagnosis is via a spirometry test. While the test can lead to a proper diagnosis, it requires 1. That the patient visits their primary care physician (PCP) on a regular basis, and 2. That the provider bothers to administer the test.

As mentioned in the introduction, individuals over the age of 65 will typically have more than one chronic condition, comorbidity. With this is mind, we propose the development of a deep learning algorithm that can predict patients who may exhibit COPD symptoms, even before a spirometry test is administered.

As previously mentioned, the modeling technique used was a feed forward neural network.. Similar to the models used to predict the other diseases, the model used for COPD prediction was a multilayer DNN design. Initial tests were made using a test harness identical to that used for diabetes prediction, as there were hopes of not having to resort to 4 separate models after all. The initial network consisted of 5 layers, the “widest” layer consisting of 600 neurons.

At first glance (figure 10), the results looked promising (PPV = 100%). Upon closer inspection though, it can be seen that while the model produced 7 true positive results, the vast majority of the results appear in the false negative and true negative categories. Thus, the initial model appears to be predicting negative diagnosis much better than positive outcomes, which is of little use to us and indicates poor sensitivity.

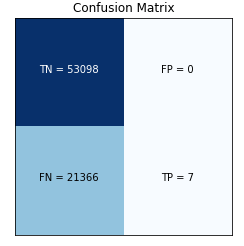


Figure 10: Initial COPD Confusion Matrix (CM), 0.9 cut-point

Given these results, it was thought that perhaps the true positive cut-point for COPD might need to be different than that for diabetes (0.9). It became clear that changing the cut-point merely changed the distribution of the totals (figure 11), and made it obvious that changes needed to be made in the model itself.

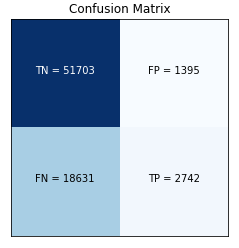


Figure 11: Initial COPD CM, 0.75 cut-point

The first step was to simplify the model such that it would run more rapidly, thus allowing less time between iterations. It was decided that the input layer could easily be reduced to the number of input nodes (in this case 82) and the neurons in the hidden layers were one by one reduced by approximately half of the initial layers, e.g., 41, 20, 10. As hoped, this resulted allowed the model to run more quickly and did not change the results in any way.

Due to the imbalance of diagnoses vs. no diagnoses, the dataset was scoured to find and remove any rows and/or columns whose values were all zero. This resulted in the removal of 10,365 rows and 1 column. Additionally, regularization logic, as tested on the other conditions, was added to each of the layers to help prevent overfitting. Finally, learning rate was decreased from 0.01 to 0.001 to ensure that we did not miss any local minima, batch size was increased to 4096, and epochs were increased to 400. These changes produced the following results (figure 12).

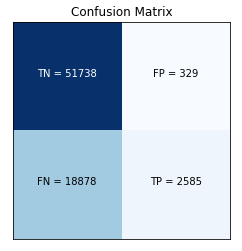
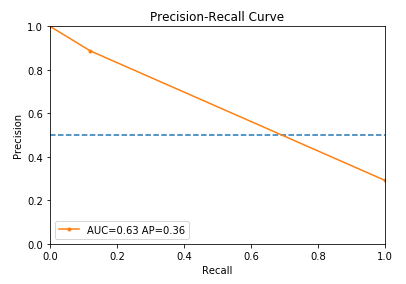
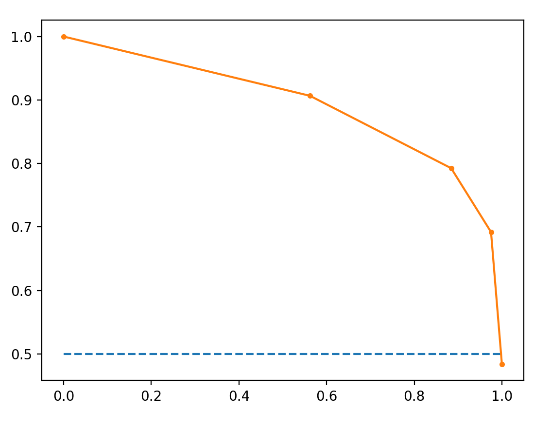


Figure 12: COPD CM, 0.75 cut-point

While we see a slight decrease in the true positive predictions, in comparison with the previous model, we also see a dramatic decrease in the false positives resulting in a PPV of 89%. While this is comparable to the diabetes model, it must still be pointed out that the cut-point for this model (0.75) is not the same as that for diabetes (0.9). The model results also show a large number of false negatives, leading us to pursue additional metrics that will help us to better tune our model. The differences between the precision/recall curves for this model (figure 13) and the ideal shape (figure 14) indicate there is a great deal of opportunity to improve.



*Figure 7: PR Curve for Current Model*



*Figure 8: Line Plot of Precision-Recall Curve*

IV. Conclusion

Similar results in this testing were obtained for Morbid Obesity, Neuropathy, and COPD: namely, failure to predict well or significantly beyond baseline probability for the condition. On the flipside, strong prediction results were obtained for Diabetes with Complications. Our takeaway from this is that the dataset and modeling methodology was suited well to such conditions as diabetes, which correlate strongly and are often cause or are caused by comorbidities. Given enough model refinement, there is possibility of usefulness for such a modeling tool to reinforce or qualify medical diagnosis and decision making with such methodology.

However, it is seen here that the methods do not fit all, or possibly even most, conditions, given this particular set of inputs. In situations where there are simpler, less expensive, testing, such as a simple BMI for obesity, such methodology is preferred. For other conditions, it may be a case of simply not having the correct inputs to predict the condition reliably or well. This leaves a vast amount of opportunity to further explore and tune models for a large volume of extant conditions.

References

1. Amadeo, K. (2018). *See for Yourself If Obamacare Increased Health Care Costs*. [online] The Balance. Available at: <https://www.thebalance.com/causes-of-rising-healthcare-costs-4064878>
2. Who.int. (2018). *WHO | Diabetes mellitus*. [online] Available at: http://www.who.int/mediacentre/factsheets/fs138/en/
3. Sharma, A. (2018). Understanding Activation Functions in Neural Networks. [online] Medium. Available at: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0
4. Keras.io. (2018). Regularizers - Keras Documentation. [online] Available at: <https://keras.io/regularizers/>
5. JAIN, S. (2018). An Overview of Regularization Techniques in Deep Learning (with Python code). [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/.
6. De Groot LJ, Chrousos G, Dungan K, et al., editors. South Dartmouth (MA): MDText.com, Inc.; 2000-.
7. Centers for Disease Control and Prevention. National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity, and Obesity. Data, Trend and Maps [online]. URL: https://www.cdc.gov/nccdphp/dnpao/data-trends-maps/index.html.
8. "Chronic Obstructive Pulmonary Disease (COPD)," [Online]. Available: https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/copd/..
9. "How and When to Use ROC Curves and Precision-Recall Curves for Classification in Python," [Online]. Available: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/.

##### Age-adjusted Hospital Discharge Rates For Neuropathy As First-listed Diagnosis Per 1,000 Diabetic Population, By Sex, United States, 1988–2007 https://www.cdc.gov/diabetes/statistics/hosplea/diabetes\_complications/fig4\_neuro.htm

##### Types Of Peripheral Neuropathy & Risk Factors | The Foundation For Pn https://www.foundationforpn.org/what-is-peripheral-neuropathy/types-risk-factors/