Global outbreak of Hepatitis Delta Virus (HDV)

Submitted by: Dayanidhi (u1189358), Christopher (u0586157), Himanshi (u1168592)

1. Problem Definition and Motivation

Healthcare is one of the dominant industries where several studies and research is being done to help improve health and prevent diseases/outbreaks. In the current generation, where colossal healthcare data is being captured every day around the world, it is very important to make a model out of the data which makes sense for us to understand how diseases are spread and what causes an outbreak.

Professor Melodie Weller is currently tracking a global outbreak of Hepatitis Delta Virus (HDV). It is estimated that 15-20 million people worldwide are infected with HDV. While this virus has been studied extensively since its discovery in the 1970s, very limited information is available to track changes in the global incidence and/or dispersed outbreaks of HDV. We planned on building upon the current global network of HDV information that is provided in public use files (PUF) from the Centers for Medicare and Medicaid Services (CMS).

Our goal was to analyze patterns in the data to track changes to HDV and HBV incidences and to identify key correlations with other diagnoses to characterize a shift in viral transmission pathways recently observed. Data obtained from the project will ultimately be made publicly available and provided as a tool for other scientists in the infectious disease epidemiology research community. If we are successful, our results may help researchers better understand the current risk factors for developing HDV, which could aid in prevention efforts. The focus of the analysis is United States of America.

2. Key Idea

HDV is a sub viral infection which grows in presence of some other virus, it is known that HBV is the major contributor of the HDV propagation. However, we wanted to identify potential causes other than HBV virus that might have supplemented the propagation of HDV and ultimately contributed to the outbreak of HDV in the United States.

To identify potential co-infecting agents of HDV, we wanted to take a large dataset of medical information, and analyze which diseases frequently appeared alongside HDV or what diseases supplements the growth of HDV virus.

Our data mining efforts were focused on above mentioned idea.

3. Data Collection and Description

To this end, we gathered datasets from two different sources that comprises of information of medical visits, diagnoses, and insurance claims of a patient. These datasets were located for us by Professor Weller, in the U of U's School of Medicine.

Dataset	**Size (in # of Rows)**	
CMS	Outpatient	14 million
	Inpatient	1.3 million
NCHS	NAMCS	1.3 million
	NHAMCS	717,000

Table 1: Table showing the dataset size

The first group of datasets came from the Centers for Medicare and Medicaid Services(CMS). It contains data regarding individual medical visits/diagnoses in the 2008-2010 period. In total, they contain ~119 million records. The upside to this group is that multiple records for the same individual carry the same ID, so we could identify a little bit of medical history per patient. The downside is that it does not cover many years, so the across-time aspect is limited.

The second group of datasets came from the CDC's National Center for Health Statistics(NCHS). Both sets are focused around cases involving Ambulatory Care, also known as Outpatient Care. Together, they have a total of about 2.42 million records, spanning the 1972-2015 period. While these records span a much larger time frame, they don't maintain a unique identifier per patient, so there is no across-time aspect in that regard.

The NCHS data provided a challenge, because it was stored in a custom data format which had to be decoded. After we took the time to decode it, we found that there was only a single record of a patient with HDV. This dataset was not going to be useful for our purposes. Hence, it was excluded from the analysis.

The CMS data, on the other hand, was stored plainly in a CSV file. But we needed to process the data to make it usable for our analysis. First, by grouping corresponding events by patient ID, so we had a full idea of each patient's medical history. Then, by converting categorical variables to a binary format. While processing this dataset, we found almost 3500 records of patients with HDV.

4. Implementation

4.1 Exploratory data analysis

To begin with exploring the CMS and NCHS data, since the dataset were huge, we decided to limit the dataset to samples that captured HDV/HBV cases for detecting the outbreak. Here are the statistics we observed in CMS dataset regarding number of HDV/HBV cases[1] -

[1] **Note:** CMS Inpatient data and NAMCS/NHAMCS didn't provide any useful results, hence they are excluded from this report.

Table 2: Table showing the # of cases of HBV or HDV or both in USA (2008-2010)

Outpatient CMS Data Samples 1-20, for the year 2008-2010			
Samples	HepB+ / HepD-	HepB+ / HepD+	HepB-/ HepD+
1	276	125	55
2	271	109	67
3	300	116	52
4	287	111	70
5	299	126	62
6	286	136	51
7	270	109	60
8	275	123	58
9	277	109	56
10	287	81	55
11	315	110	50
12	275	103	71
13	255	107	56
14	309	130	67
15	292	129	49
16	279	127	52
17	284	132	52
18	277	134	53
19	278	124	65
20	299	136	44
Total	5691	2377	1145

Handling duplicates:

Total	**5646**	**2360**	**1138**

Table 3:HDV percentages with respect to HepB

{+D/+B} %	{-B, +D}/+D %
0.436539415	0.3253287593

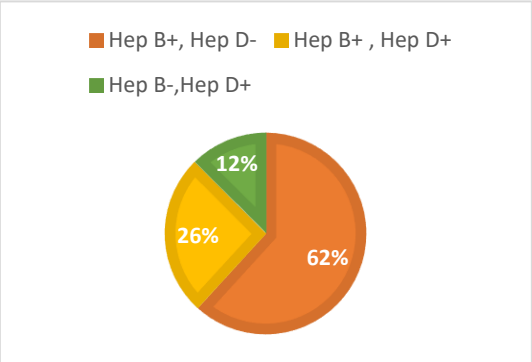


Figure 1: Percentage distribution of HDV/HBV cases in the USA (2008-2010)

We have compared our statistics that we generated on our datasets with the statistics that are published in the research community. For the percentage of Hep B patients that had HDV, our stats showed around 44%, whereas the industry claims 5%[2]. Part of the problem was there were duplicates in the dataset (where a patient had multiple visits and there were multiple records for same patient), but this turns out to be a small factor. The major factor influencing this percentage difference is likely a bias in the sample population. There may be high risk patients, who are more likely to contract multiple diseases.

Along with that we also investigated the spread of these cases (HDV/HBV) across the USA, to identify the high-density area vs low density area with respect to HDV/HBV and use information in predictive modeling. From this exploration we found that coastal areas are highly affected by HBV/HDV viruses.

[2] Source: http://www.who.int/mediacentre/factsheets/hepatitis-d/en/

Hepatitis Delta cases in USA (2008-2010)

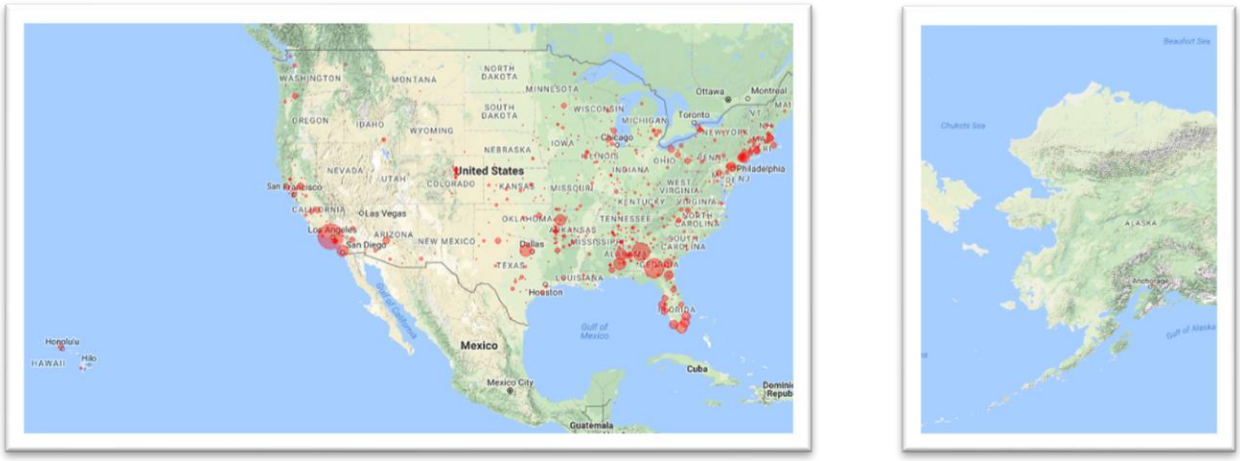


Figure 2:Hep D cases in USA (2008-2010) grouped by county (On right- Alaska)

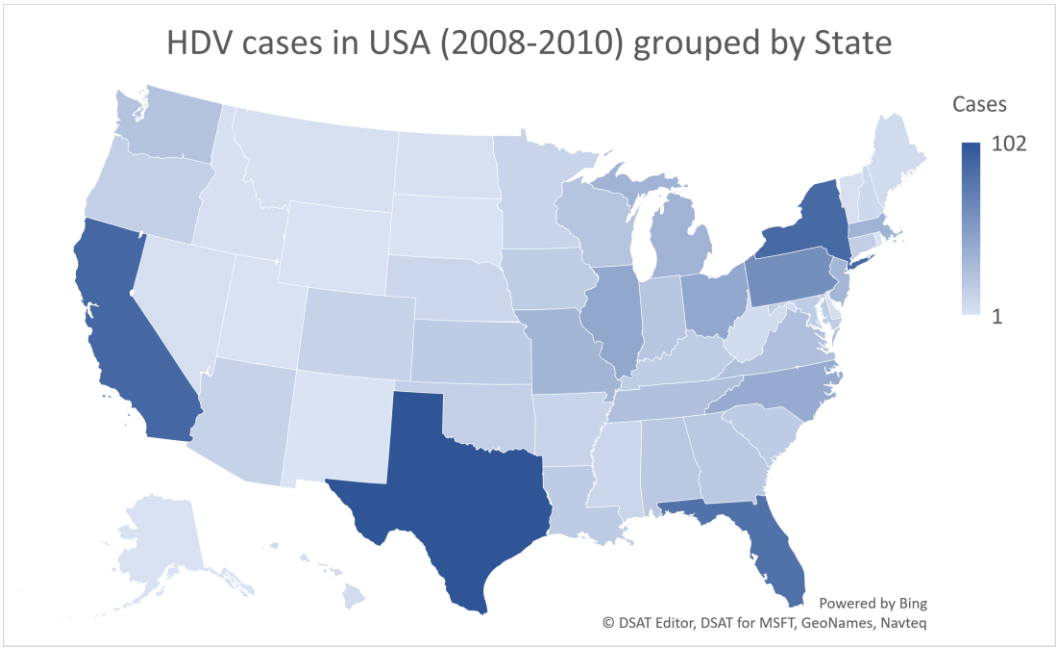


Figure 3: Hep D cases in USA (2008-2010) grouped by State

Hepatitis Delta existence with Hepatitis B cases in USA (2008-2010)

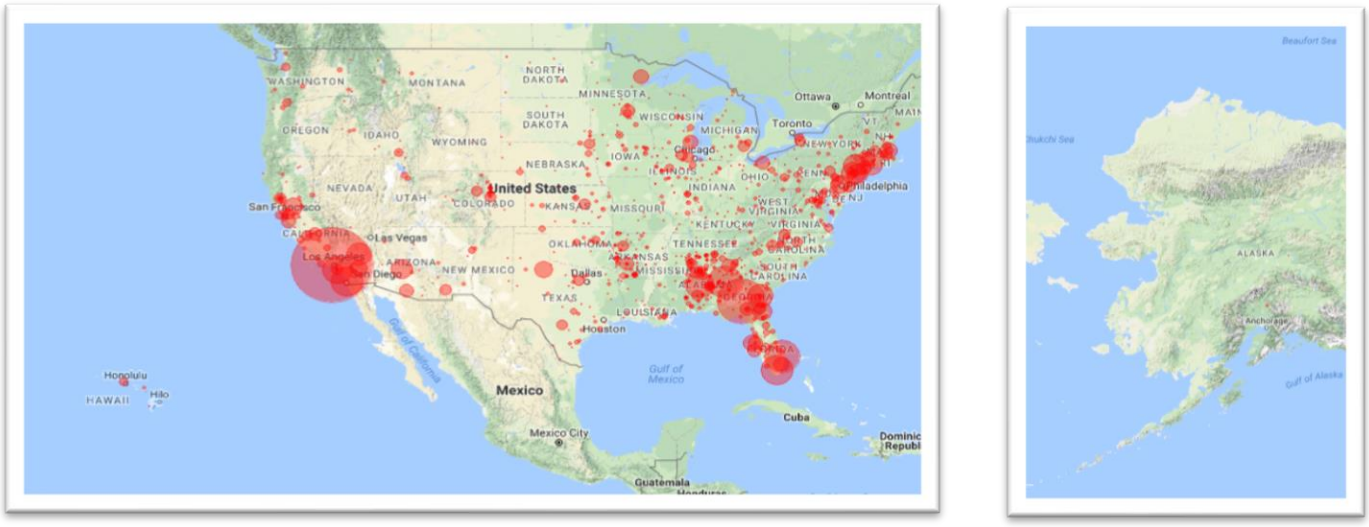


Figure 4: HBV co-occurrence HDV positive cases in USA (2008-2010) grouped by county (On Right – Alaska)

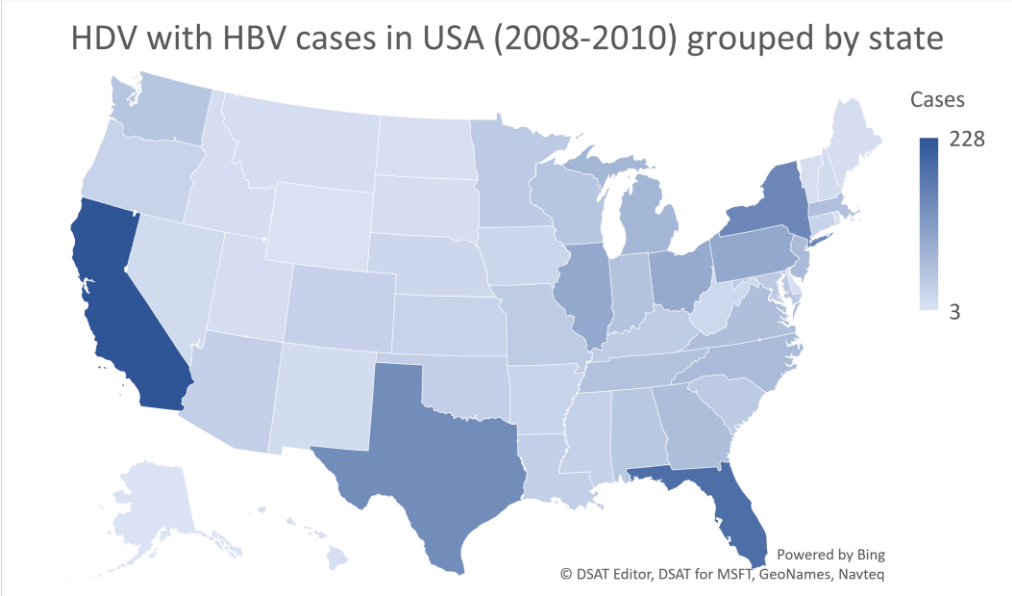


Figure 5: HBV co-occurrence HDV positive cases in USA (2008-2010) grouped by State

Hepatitis B cases in USA (2008-2010)

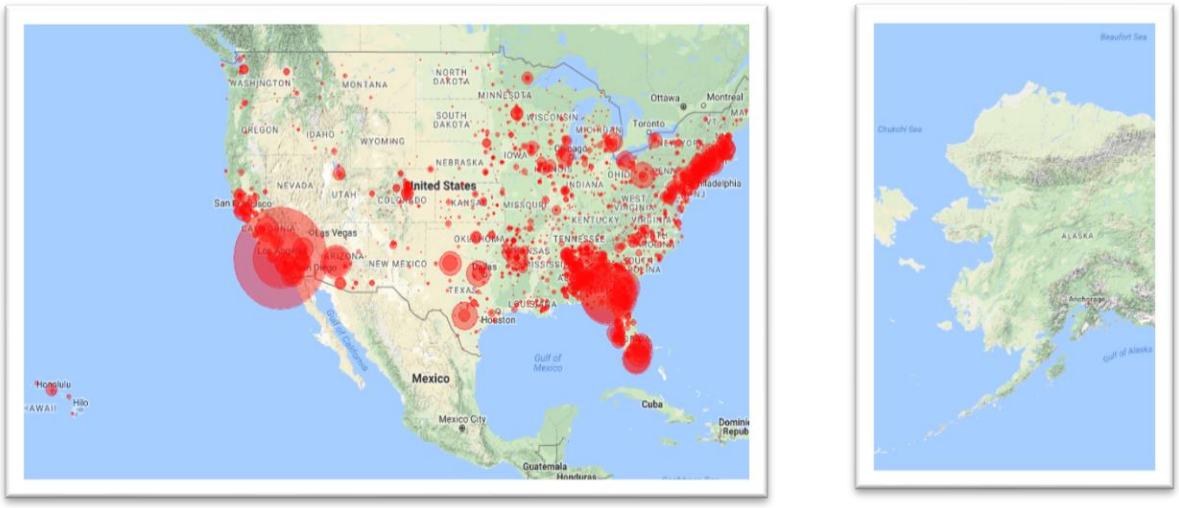


Figure 6: Hepatitis B cases in USA (2008-2010) grouped by county (On Right – Alaska)

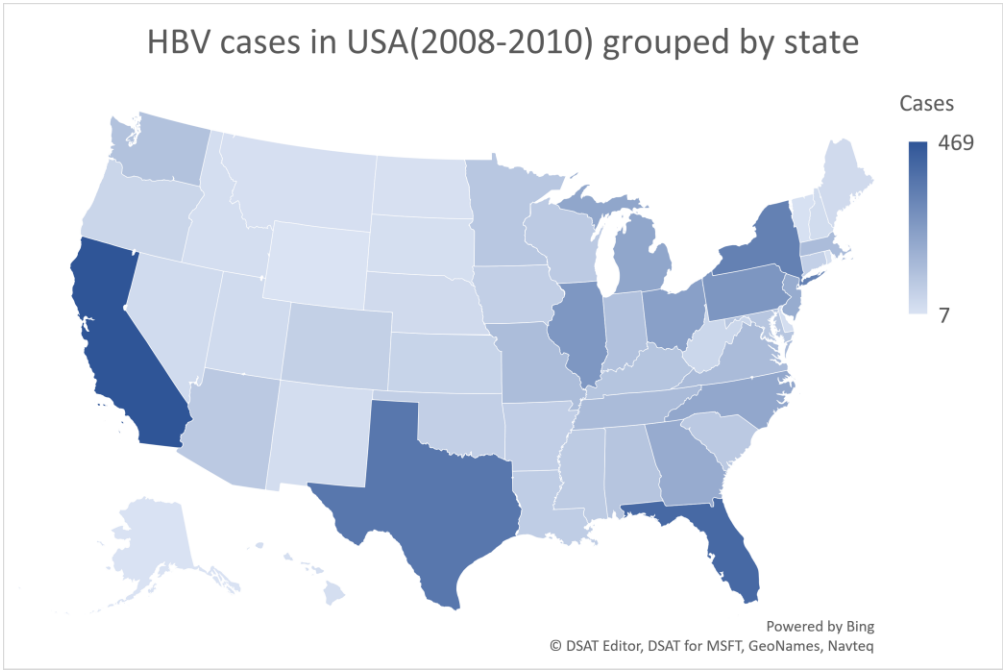


Figure 7: Hepatitis B cases in USA (2008-2010) grouped by State

Please follow link for an interactive graph, which shows individual county total cases:
https://himanshi-27.github.io/

Now, to analyze which diseases, co-occur with HDV, we started off by applying a Misra-Gries/Frequent Item datamining technique as our preliminary search to see which two diagnosis codes occur together frequently, to get an inkling of what other diseases co-exists with HDV other than HBV.
Here are the observations –

Table 4: Table showing Misra-Gries observation on our pre-processed data.

Top 24 Frequent items, Using Misra-Gries Algorithm (k=100)		
S.No	Diagnosis Code Pairs	Frequency Counter
1	2724_4019	74
2	7030_70715	6
3	5715_71590	4
4	7030_78720	9
5	7032_25000	104
6	7030_41400	36
7	2809_7030	181
8	311_7032	41
9	3051_7032	35
10	2720_4019	37
11	7032_78720	10
12	7030_53012	2
13	7030_42731	56
14	5715_25000	31
15	2859_7906	10
16	7030_38812	2
17	4019_7030	334
18	4019_7032	212
19	4011_25000	32
20	4019_7840	10
21	2809_27542	4
22	7030_25000	152
23	2721_7032	10
24	2721_7032	10

Challenges: In this technique of Frequent items, using a small 'k' such as 10 isn't useful for such a big dataset, as it always resulted in the last few items being frequent items regardless. Increasing it to 100 made more sense as it gave the necessary scope for frequent items to stay in the queue much longer.

Since HBV, HDV cases for a patient might not both be in the sample dataset, it presented an interesting problem. We had to consider HDV+ meant HBV+ as well if data isn't available.

We have compared our statistics that we observed on our datasets with the statistics that were published by CMS. For analyzing the frequent items results, we used bootstrapping to get an estimate of how common these other diseases are in the overall dataset, and that will tell us the significance of their frequency amongst HBV/HDV patients.

After doing Frequent Items, we also tried using Bootstrapping to accomplish our goal. When looking at the diseases which often occur with HDV, it can be hard to tell whether they're frequent just because they're common diseases or if they're more frequent than normal. We can get an idea of this by comparing their frequency with their frequency in the overall dataset, but again, it's hard to tell when a difference is significant.

By randomly resampling from the population, we can estimate the distribution of these frequencies. Through many random samples, we can see how often a given disease appears to have a certain proportion. By comparing our HDV proportions with these distributions, we can get an idea of how uncommon our findings are.

After running this analysis, we narrowed our ~15,000 diagnoses down to ~3,700 diagnoses that somewhat stand out, and ~580 diagnoses which exceptionally stand out, as these codes co-occur with HDV to a significantly different degree than the general population.

4.2 Predictive modeling

To test bootstraping result, we decided to use Logistic Regression to evaluate whether these 580 diagnosis codes were truly significant indicators of HDV. We reduced our dataset to only contain these codes and tried to predict the odds of a patient having HDV.

Input: Dataset with column: Age, state, county, dummy variables of ~580 exceptional codes (0: don't have the disease, 1: have diagnosed with the disease), other medical information curated from beneficiary summary file of CMS dataset, along with dummy variable HBV. The target variable was HDV, which has value 0 or 1, where 1 represents HDV+ cases. (1703392 observations against 858 variables)

A model was also build with considering HBV virus, however the AIC value for this model was very large as compared to the model with HBV, so we rejected the model built without HBV virus.

Output: Odds of having HDV.

Challenges: We were faced with two challenges:

1. The dimensions were a lot.
2. The imbalance in classes, there were a lot non HDV cases.

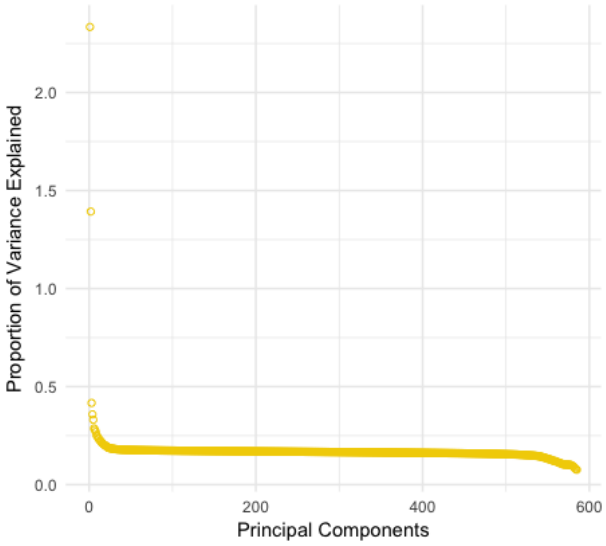


Figure 8: Percentage proportion variation explained by principal components

To tackle these challenges, we first utilized Principal Component Analysis(PCA) to reduce the dimensions, however, the results of PCA were not strong as the First Principal Component only explained about 2% variation in dataset, below is plot showing the percentage variation explained by components.

Given the results of PCA, we decided to move ahead without reducing the dimensions.

For the next challenge, we wanted to balance the classes because our logistic classifier holds a bias towards the majority class tends to classify majority class more often. We wanted our classifier to classify both classes without any bias for either of the classes. We used a method of Synthetic Minority Over-Sampling Technique (SMOTE)[3] to balance the classes. This sampling method uses the combination of over-sampling and under-sampling technique for balancing classes, in this method we over-samples the minority class and under-samples the majority class. We have doubled the minority classes and halved the majority class.

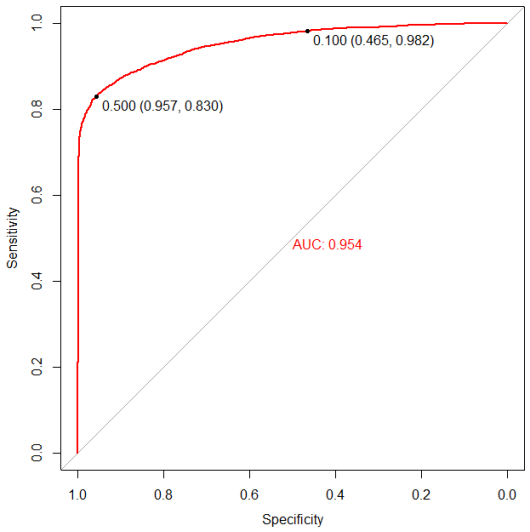
	No (HDV – cases)	**Yes (HDV+ cases)**
Original	1359926	2788
After resampling with SMOTE	5576	5576

Table 5: Table showing the original class distribution and resampled distribution

Post data preparation we build a model with HDV as target variable and other variables as the predictive variables. The training and testing set was created as an 80-20 split of whole data set respectively.

We found that HBV is highly significant in determining the odds of having HDV, which matches the known fact about HDV infection. Moreover, we also found that critical conditions such as Diabetes, Heart diseases,

[3] Source: https://www.jair.org/media/953/live-953-2037-jair.pdf



Osteoporosis, Arthritis Osteoarthritis, and Chronic kidney disease are also significant for the predictions of odds of having HDV.

Another challenge post modeling was to decide the threshold to say whether this probability value represents "HDV +" cases or "HDV – cases". Since, with medical data we want less number of predictions that identify a person infected with HDV as not being identified with HDV.

We evaluated that for this dataset "0.11" is a good estimate for saying any probability greater than this are "HDV +" cases, it gives around 1% false negative classes. This result was estimated using 10-folds cross validation technique.

Following are the evaluation results on training and testing set [4], showing the confusion matrix, sensitivity and specificity evaluation on training and test data.

Figure 9: ROC curve for Logistic Model

Data Partition							
Training set	Original No	Original Yes	Reference(right)/ Predicted(down)	No	Yes	Sensitivity	Specificity
	5576	5576	No	2825	130	0.5066	0.9767
			Yes	2751	5446		
Testing set	Original No	Original Yes	Reference(right)/ Predicted(down)	No	Yes	Sensitivity	Specificity
	339978	700	No	166371	76	0.4894	0.8914
			Yes	173607	624		

Table 6: Table showing the evaluation results of logistic model on training and test set.

5. Learning outcomes

We learned that understanding data points is crucial before we interpret them for meaningful use. Diagnosis codes have a special significance for leading zeros. 07020 is different from 7020 (07020: Hepatitis B without HDV, 7020: Actinic keratosis). We considered both initially by mistake, but now used the correct codes to generate our statistics.

Overall, we found that while we were able to use Misra-Gries and Bootstrapping to find some diagnoses which tend to disproportionately appear with HDV, these codes were not great indicators of HDV. It seems that our bootstrapping test alone is not sufficient for determining significance.

[4] The sensitivity and specificity are evaluated considering "No" as the positive class.

Appendix A:

We have used CMS provided public data:

https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html

https://www.resdac.org/cms-data/files/de-synpuf

The Codebook for understanding CMS data:
https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_Codebook.pdf

About HDV in Sjögren's Syndrome Patients:
https://www.ncbi.nlm.nih.gov/pubmed/27294212

Appendix B:

Work item	Person(s) who did work on this
Downloading different data sets and combining those data sets.	All three of us. There were 4 different file sets. Chris Peterson Himanshi Sharma Dayanidhi Tandra
ICD9 code 07020 is different from 7020 (07020: Hepatitis B without HDV, 7020: Actinic keratosis). There were other such diagnosis codes where we had to differentiate these codes based on leading zero.	All three of us. There were 4 different file sets. Chris Peterson Himanshi Sharma Dayanidhi Tandra
There were multiple visit information for same patient (multiple rows for same patient). We had to combine them to have one row per patient.	All three of us. There were 4 different file sets. Chris Peterson Himanshi Sharma Dayanidhi Tandra
Calculate # of cases of HBV or HDV or both in USA (2008-2010) for CMS dataset	Himanshi Sharma Dayanidhi Tandra
Calculate # of cases of HBV or HDV or both in USA for NAMCS/NHAMCS	Chris Peterson
Frequent Items using Misra-Gries to identify which ICD9 codes are occurring together.	Dayanidhi Tandra
Creating HDV/HEP percentages by treating multiple visits of a patient as one line item.	Chris Peterson Himanshi Sharma Dayanidhi Tandra
Boot Strapping	Chris Peterson
Creating Abstract datatype (Matrix) from the datasets.	Chris Peterson Himanshi Sharma Dayanidhi Tandra
Logistic Regression	Himanshi Sharma