Modeling an Annotated Database on Autism

Jiashan Liang, *MIT EECS Undergraduate Research and Innovation Scholar*, Peter Szolovits, *MIT*, and Finale Doshi-Velez, *Harvard Medical School*

Abstract—**Autism Spectrum Disorder (ASD) is identified in about 1 in 68 US children, and there are many common co-morbidities that occur with ASD. Characterization of the disease still requires work and we aim to better understand the co-occurence patterns of co-morbidities by collecting an annotated database on children with ASD. This annotated database will create a scaleable and reproducible method to characterize ASD patients. This paper focuses on discovering the co-morbidities that occur with ASD by running hierarchical clustering to determine the groups of similar patients. Previous related work identified clusters using codified data from electronic health records. We build upon that work by using a combination of codified data and data from narrative text obtained by using Natural Language Processing (NLP). The narrative text includes information from doctors' and nurses' notes and discharge summaries. We compare the resulting clusters with those found using only codified data. Running clustering with both codified and NLP data resulted in one cluster. We saw a multi-cluster result when using only the codified data. We conclude that the NLP data needs to be filtered further to reveal underlying relationships among ASD patients.**

Index Terms—**Autism, Comorbidity, Clustering**

I. Introduction

THE OVERALL goal of the project is to create an annotated database on children with Autism Spectrum Disorder (ASD). This disease affects 1% of the population [4], and is associated with a variety of co-morbidities, for example gastrointestinal and seizure disorders. Even though the disease is prevalent, its characterization still requires work. Creating and analyzing an annotated database on children with ASD could result in a scaleable and reproducible method to characterize ASD patients. The database will allow us to better understand the co-occurrence patterns of co-morbidities.

Previous work has been done to quantify the occurrence of various co-morbidities in ASD, including gastrointestinal disorders [1] and sleep disorders [6]. These projects have used a limited number of patients and most projects have focused on the relationship between ASD and one specific disorder [2].

This paper focuses on the first part of the project, which is to discover relationships among ASD patients. We currently have de-identified data on about 4,000 autistic patients from Boston Children's Hospital (BCH). The data is drawn from two main sources to better explore the relationship between ASD and clinical manifestations beyond neurobehavioral criteria. The first is codified data, which consists of lab studies, medications, and diagnoses. The second is narrative text, which includes things such as doctors' and nurses' notes and discharge summaries. Doshi-Velez has done basic clustering on a set of codified data, but our use of the narrative text is novel [2]. We build upon her work to run hierarchical clustering on different data and try to identify interesting groups of similar patients.

For the next part of the project, work will be done to extract patient information from clusters so that the clusters can be characterized by the clinical state of the patients belonging to each cluster. We anticipate getting similar data on about 15,000 patients and possibly a similar number of controls. Controls in this case are patients, but ones who are not characterized as having ASD. These models will enable us to relate a child's data to other diseases from which they might suffer.

II. Previous Work

There has been a variety of previous research that revealed relationships between ASD and other diseases. For example, Horvath received survey responses from 112 children and concluded that 76% of ASD patients had at least one gastrointestinal symptom, whereas only 30% of non-ASD patients had one gastrointestinal symptom [1]. Similar studies have been done to show a relationship between ASD patients and epilepsy [5], sleep disorders [6], and muscular dystrophy [7]. All these papers used less than 200 patients as the sample size due to the limitation of using surveys and diagnostic tests. In addition, these papers were examine the relationship between ASD and one specific disorder [2].

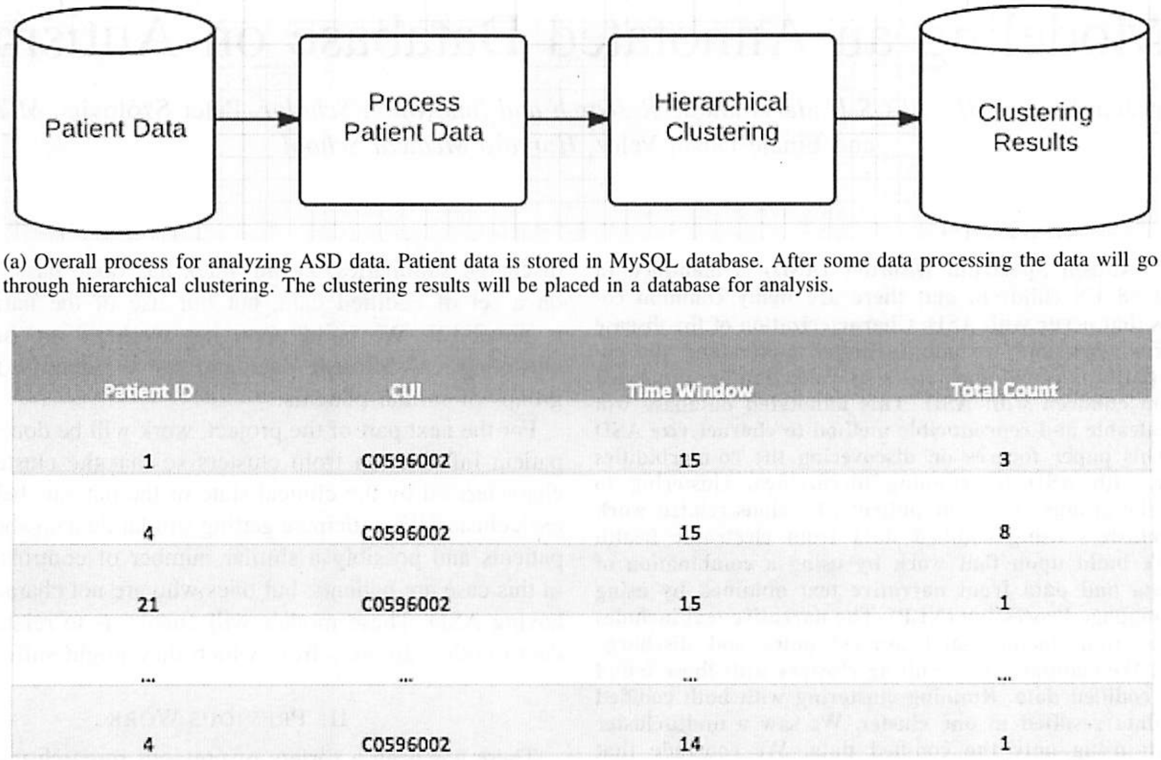
Doshi-Velez examined the patterns of co-morbidities in 4927 ASD patients by using the International Classification of Diseases, Ninth Revision (ICD-9) codes from electronic medical records. She used hierarchical clustering to find that the patients were clustered into four major groups. One group was characterized by seizures, another by multi-system disorders including gastrointestinal disorders and auditory disorders, and the third by psychiatric disorders. The fourth group could not be further resolved.

III. Process

We want to filter the patient data and use hierarchical clustering to find clusters. In this section, we cover the methods used to cluster ASD patients, along with a description of the patients.

A. Overview

Figure 1a is a visual representation of the overall structure of this project. Data processing is first done to combine the codified data (ICD-9 codes) and narrative text. Natural language processing (NLP) has been used to extract information from the narrative texts. Specifically, the text has been analyzed



(a) Overall process for analyzing ASD data. Patient data is stored in MySQL database. After some data processing the data will go through hierarchical clustering. The clustering results will be placed in a database for analysis.

Patient ID	CUI	Time Window	Total Count
1	C0596002	15	3
4	C0596002	15	8
21	C0596002	15	1
...	...	...	...
4	C0596002	14	1

(b) An example table. The second row shows that Patient 4 had a CUI=C0596002 three times when he was 15 years old (time window=15). He had CUI=C0596002 once when he was 14 years old.

Fig. 1. Description of overall project. Patient data is filtered and then hierarchical clustering is used to store find similar groups of patients.

using clinical Text Analysis and Knowledge Extraction System (cTAKES), an open-source NLP system that takes clinical notes and identifies types of clinical named entities.

The results from cTakes and codified data are stored in a MySQL database. Each entry corresponds to a patient visit, and the columns contain visit time and associated information during that visit. The relevant data includes Concept Unique Identifiers (CUIs), blood test results, drug dosages, etc. CUIs are used by the United Medical Language System (UMLS) to denote unique concepts from different medical sources. For example, a CUI of C0018681 refers to headaches. We are interested in seeing how patients can be grouped together by CUIs.

After the data processing, we find relationships among patients by running clustering. In the future, work will be done to extract the patient information from the clusters. The results from clustering will be put into a database and we will create a program that will identify the CUIs that describe each cluster.

B. Patients

Patient data, including ICD-9 codes and NLP data, were provided by the Boston Children's Hospital. After filtering for male and female patients who are at least 15 years old, we had 3654 patients that we were using for clustering. We are also planning on receiving data on 15,000 patients that will be provided by the BCH for future clustering analysis.

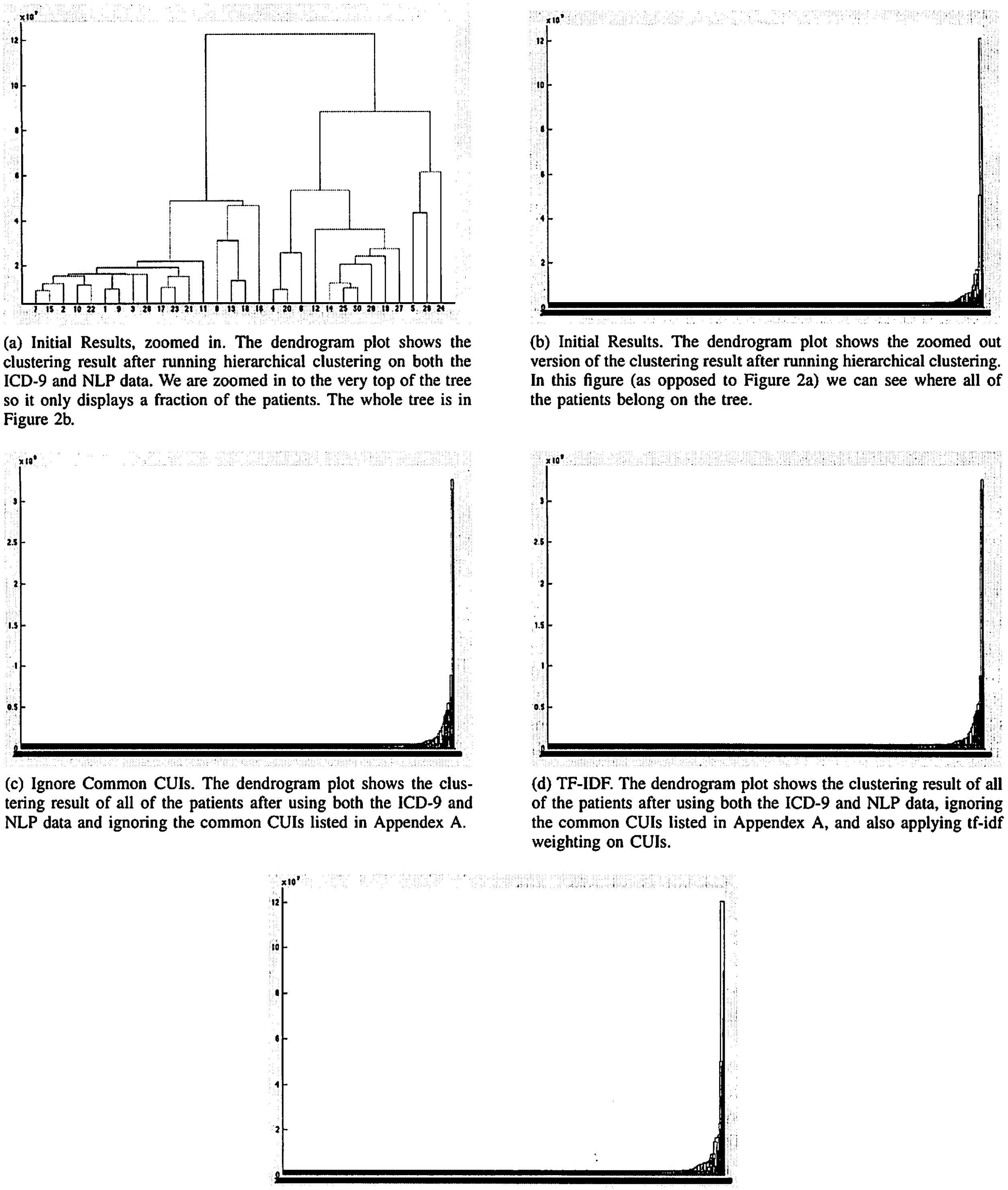
C. Methods

We performed some similar data processing as Doshi-Velez did [2]. She created 30 six-month windows from a patient's birth to age 15. From each time window, she calculated the number of occurrences of certain categories for each patient. By using hierarchical clustering with Euclidean distance, she found four major clusters. We have added NLP data to this process to see how it impacts the clustering.

We created a time-series with 15 1-year windows from birth to age 15. Looking at each time window table and determining how many times a patient had a CUI, we then constructed a table for the CUIs found in the ICD-9 data. We did the same thing for the CUIs we received in the NLP data. We then combined the ICD-9 table and NLP table. Figure 1b shows some example table entries.

After combining the tables, we performed k-Nearest Neighbors (k-NN). We calculated the distance squared between patients by seeing how many CUIs they shared in each time window. We determined each patient's k=10 closest patients and stored the information in a 3654x3654 table that has the distances from any patient to its 10 closest patients.

We performed hierarchical clustering using Euclidean distance and Ward's method on the k-NN table to find clusters of similar patients. (See Appendix A for a more detailed description of hierarchical clustering). Matlab was used to calculate the hierarchical cluster tree. We examined the dendrogram plot of the cluster tree to see how many clusters were formed.



(a) Initial Results, zoomed in. The dendrogram plot shows the clustering result after running hierarchical clustering on both the ICD-9 and NLP data. We are zoomed in to the very top of the tree so it only displays a fraction of the patients. The whole tree is in Figure 2b.

(b) Initial Results. The dendrogram plot shows the zoomed out version of the clustering result after running hierarchical clustering. In this figure (as opposed to Figure 2a) we can see where all of the patients belong on the tree.

(c) Ignore Common CUIs. The dendrogram plot shows the clustering result of all of the patients after using both the ICD-9 and NLP data and ignoring the common CUIs listed in Appendex A.

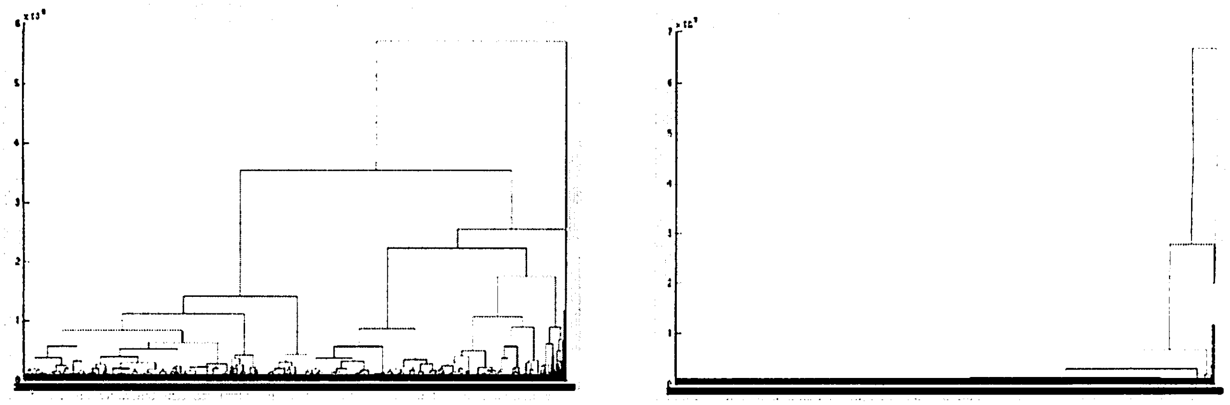
(d) TF-IDF. The dendrogram plot shows the clustering result of all of the patients after using both the ICD-9 and NLP data, ignoring the common CUIs listed in Appendex A, and also applying tf-idf weighting on CUIs.

(e) NLP data by itself. The dendrogram plot shows the clustering result of all of the patients after using only the NLP data and not including the ICD-9 data.

Fig. 2. Hierarchical clustering results. These figures display the dendrogram plot of various clustering results. The patients are on the x-axis, and they are getting clustered together as we go up the y-axis. We can see that in all runs, the patients are being clustered in one group.

IV. Results

We are interested in finding the co-morbidities among ASD patients. After processing the patient data, we used hierarchical clustering to identify major clusters of patients. We ran the process described in the Method section eight different ways, with slight changes during each run as described below. For almost every run, we found that all the patients were grouped into one cluster. The only time we saw multi-cluster results was when we only used ICD-9 patient data. A summary of



(a) ICD-9 data by itself with no k-NN clustering. The dendrogram plot shows the clustering result of all of the patients after using only the ICD-9 data and not using k-NN clustering. It is the only dendrogram plot with two major clusters.

(b) NLP and ICD-9 data with no k-NN clustering. The dendrogram plot shows the clustering result of all patients after using both the NLP and ICD-9 data and not using k-NN clustering.

Fig. 3. Hierarchical clustering results with no k-NN clustering. The first figure displays the dendrogram plot of running with only ICD-9 data and no k-NN Clustering. This is the only dendrogram with

the runs can be found in Table 1. Figures 2 and 3 show the corresponding dendrograms for the different runs.

A. *Initial Results*

In the initial run, we combined the ICD-9 and NLP patient data and ran k-NN and hierarchical clustering. As we can see in the dendrogram plot shown in Figure 2a and 2b, the patients are mostly grouped into one cluster. We noticed that some CUIs appear among a significant portion of patients. We hypothesized that the amount of CUIs could be hiding the differences among patients and causing the hierarchical clustering to group all the patients together.

B. *Ignore Common CUIs*

As a result of the initial run, we decided to remove some of the common CUIs that many patients have. We looked at the top 200 most commonly occurring CUIs and selected 34 CUIs that do not seem to contribute much to our calculations. For example C0262926, which refers to Medical history, was removed because almost all patients have a medical history so this CUI would not help differentiate between patients. A full list of the removed CUIs can be found in Appendix B.

After the 34 CUIs were added to an ignore list, we reran the experiment with a slight modification. When comparing how similar patients were during the k-NN step, we first checked if each CUI was in the Ignore list. If it was, we did not factor that CUI in when determining how similar two patients were.

Figure 2c shows the clustering result after removing the common CUIs. As we can see, the left tail, single cluster pattern appeared again. The dendrogram looks almost identical to the one shown in Figure 2b. Because we only removed 34 CUIs, we decided to try to lower the impact of common CUIs even more.

C. *TF-IDF Weighting*

Since removing the most common CUIs resulted in the same one cluster phenomenon, we decided to further lessen the effects of common CUIs by applying term frequency-inverse document frequency (tf-idf) to weight the patient CUIs. (See Appendix C for more information about our implementation of tf-idf). By weighting each CUI with tf-idf before we performed k-NN, we wanted to lower the impact of commonly seen CUIs and thus reveal more than one cluster.

Figure 2d shows the clustering result after applying tf-idf. The single cluster still exists, which is an interesting result because tf-idf weighting and ignoring the 34 common CUIs should have helped remove any hidden similarities hidden by the common CUIs.

D. *Analyze Only NLP Data*

In the three previous sections, we were combining the ICD-9 data with the NLP data. In this run, we try using only the NLP data by itself and ignoring the common 34 CUIs and using tf-idf weighting. The resulting clustering is shown in Figure 2e. It once again displays a similar one cluster pattern we saw earlier. We will examine the one cluster result more in the Discussion section.

E. *Varying Time Windows*

Using both the ICD-9 data and NLP data, we ignored the 34 common CUIs and tried varying the size of the time windows. We tried one time window and 30 six-month time windows. Regardless of time window size, we still saw the one cluster phenomenon.

F. *Analyze Only ICD-9 Data*

Since the Dosh-Velez paper only used ICD-9 data, we decided to run the experiment using only the ICD-9 data. We first aggregated the ICD-9s into phenome-wide association study (PheWAS) codes [8]. There were a total of 802 PheWAS categories. After running hierarchical clustering, the patients were still grouped into one cluster.

TABLE I
SUMMARY OF RUNS

Run	NLP Data	ICD-9 Data	Ignore Common CUIs	TF-IDF Weighting	Time Window	k-NN Clustering	Num Clusters
A	Yes	Yes	No	No	1 year	Yes	1
B	Yes	Yes	Yes	No	1 year	Yes	1
C	Yes	Yes	Yes	Yes	1 year	Yes	1
D	Yes	No	Yes	Yes	1 year	Yes	1
E	Yes	Yes	Yes	No	6-months	Yes	1
F	No	Yes	No	No	6-months	Yes	1
G	No	Yes	No	No	6-months	No	2
H	Yes	Yes	Yes	No	1-year	No	1

G. Analyze Only ICD-9 Data and No k-NN Clustering

In addition, to only analyzing the ICD-9 data, the Doshi-Velez paper did not perform k-NN clustering before running hierarchical clustering. After aggregating the ICD-9 data into PHeWAS codes, we tried going straight to clustering without running k-NN. The clustering result is shown in Figure 3a. The figure has two major clusters, which is more along the lines of what Doshi-Velez found.

What exactly does k-NN do?

H. Analyze All Data and No k-NN Clustering

Since we saw a multi-clustering result after removing k-NN clustering in the previous run, we tried using both the NLP and the ICD-9 data (as in Result's section B), and removing the k-NN clustering. However, we still got the one cluster graph as shown in Figure 3b.

V. Discussion

We are interested in the fact that clustering using only ICD-9 data showed four clusters as discussed in Doshi-Velez's paper [2], but adding the NLP data resulted in all the patients being placed in one cluster. We hypothesize that there are a lot of extraneous NLP data that are covering the results.

When we ran the experiment with only ICD-9 data and removing k-NN clustering, the patients were clustered into two groups as can be seen in Figure 3a. The clustering result, along with Doshi-Velez's results, implies that there inherently exists a relationship between patients with ASD and other diseases. However, adding the NLP data and removing k-NN clustering resulted in a one cluster result.

The table shown in Figure 1b was created from about 3 million entries from the NLP data and about 50,000 entries from the ICD-9 data. We think that the sheer volume of NLP data compared to ICD-9 data could be covering the relationships among patients. As shown in the Results section, we have already tried to reduce the noise from NLP data by removing the 34 common CUIs. However, more filtering should be done on the NLP data to reduce the noise.

Word up

Also, removing the k-NN clustering step was necessary to get the two cluster result. In other words, the hierarchical clustering was formed by looking at a patient's relationship to every other patient. We initially used k-NN clustering to find a patient's top ten most common neighbors because we thought that should be sufficient to identify major clusters and it is faster to run hierarchical clustering. More work should be done to determine if can use a smaller number of patients.

A. Challenges and Limitations

One of the restrictions on our data was that we could only received de-identified patient data from Boston Children's Hospital. We wanted to try different methods of NLP instead of cTAKES but we were restricted to cTAKES due to the fact that NLP was running on the identified patient data. Since different NLP programs will extract different CUIs, it would be beneficial to see the resulting clusters from another program.

VI. Future Work

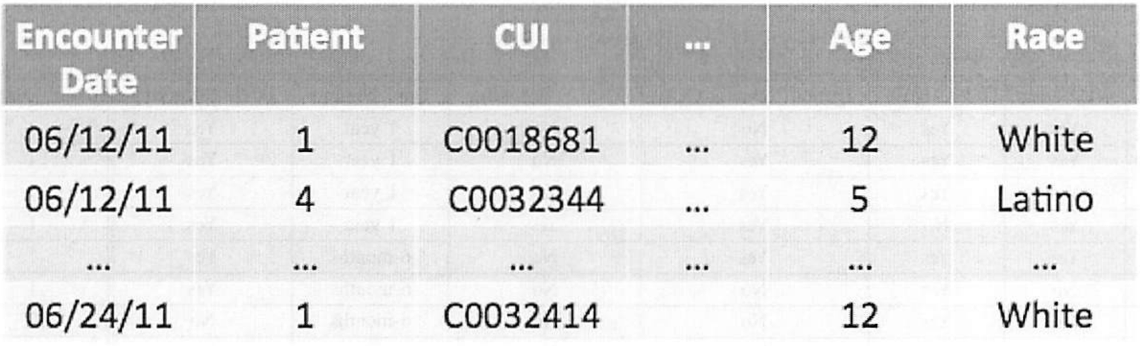
There are three types of future work we are considering. The first involves working with more data. We would like to get more patient data from the hospital to run the experiments on a larger pool of patients. We would also like to try varying the NLP performed on the patient data to see if having more data will result in better clustering.

Secondly, a variety of work could be done to further preprocess the data. Currently, there are about 4.18 million entries for 154 patients because a single patient has many entries in the database each time he visits the hospital, corresponding to every measurement and clinical description (see Figure 4a). Since we hope to get data for about 1000 times as many patients, the database will become very long. We are interested in creating a new database that contains only one entry for each patient. An entry will consist of a combined version of all of the visits that the patient has had. This new format will make it easier to pre-process the data (see Figure 4b).

Also, the patient data is coming from Boston Children's Hospital, so some column values are unclear. Since we have so much data, we are interested in removing columns that are not relevant to our project. For example, a column that stores the date that the hospital staff entered in patient information is not necessary.

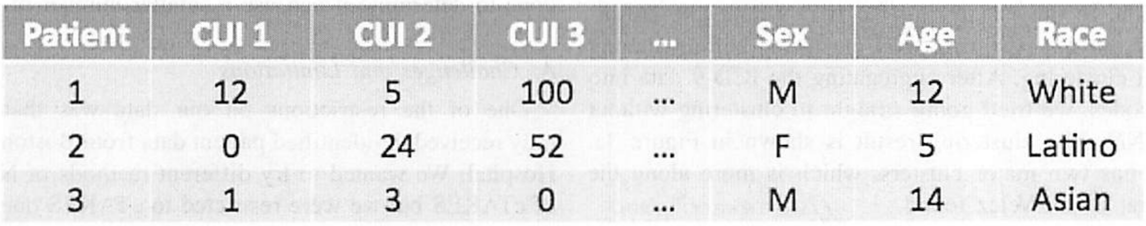
In addition, as we mentioned in Discussion section, more pre-processing should be done on the NLP data. For instance, the NLP data could be filtered first by semantic type, which is a category used by the UMLS to describe data. An example of a semantic type is "Clinical Drug", which could be associated with a patient entry.

The third aspect of future work is to proceed to the next part of the project. As described earlier, work will be done to



Encounter Date	Patient	CUI	...	Age	Race
06/12/11	1	C0018681	...	12	White
06/12/11	4	C0032344	...	5	Latino
...	...	...	...	...	...
06/24/11	1	C0032414	...	12	White

(a) Sample entries in database before data processing. Note that the same patient will appear more than once if they are admitted to the hospital more than once. CUI indicates what symptoms/problems the patient had during that visit. For example, C0018681 represents a headache.



Patient	CUI 1	CUI 2	CUI 3	...	Sex	Age	Race
1	12	5	100	...	M	12	White
2	0	24	52	...	F	5	Latino
3	1	3	0	...	M	14	Asian

(b) Sample entries in database after data processing. Note that each of the CUI entries indicates how many times that CUI showed up over all patient encounters. For example, if CUI 1 was a headache, then patient 1 had a headache 15 times over all hospital visits.

Fig. 4. Potential future work includes more data processing. The current patient data includes unnecessary information and could be streamlined more.

get patient information from the clusters found in hierarchical clustering. Each cluster will be characterized by the diseases of the patients belonging to it. But of course this depends...

VII. Conclusion

The goal of the project is to create an annotated database for patients with ASD and to use this database to identify co-occurrence patterns of co-morbidities. In the first part of the project, we used clustering on patient data to try to reveal underlying relationships between different groups of children who have autism.

After running k-NN and hierarchical clustering on the ICD-9 and NLP data of 3654 patients, we saw that all the patients were grouped in one cluster. Only running the clustering with ICD-9 data and not using k-NN clustering resulted in a two-cluster graph. We concluded that the amount of NLP data is hiding the relationships among patients and that it needs be filtered more. Future work includes replicating the experiment on more data and improving the data preprocessing.

Acknowledgments

We would like to thank John Bickel and other staff members of the Boston Children's Hospital for getting us the de-identified patient data.

References

[1] K. Horvath, J. Papadimitriou, A. Rabsztyn, C. Drachenberg and J. Tildon. (1999). *Gastrointenstinal abnormalities in children with autistic disorder.* The Journal of Pediatrics, 135:559563.

[2] F. Doshi-Velez, Y. Ge and I. Kohan. (2014). *Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Anaylsis.* Pediatrics, 133:1.

[3] P. Szolovits and I. Kohan. (2012). *Matched Cohort of Autism Spectrum Disorder and Controls from Electronic Medical Records.* Unpublished proposal.

[4] J. Baio. (2012). *Prevalence of autism spectrum disorders autism and developmental disabilities monitoring network, 14 sites, united states, 2008.* CDC Morbidity and Mortality Weekly Report, 61:1-19.

[5] SE Mouridsen, B. Rich, T. I. (1999). *Epilepsy in disintegrative psychosis and infantile autism: a long-term validation study.* Dev Med Child Neurol, 41:403411.

[6] A. Richdale, K. Schreck, T. I. (2009). *Sleep problems in autism spectrum disorders: Prevalence, nature, and possible biopsychosocial aetiologies.* Sleep Medicine Reviews, 13:110114.

[7] J. Y. Wu, K. C. K. Kuban, E. Allred, F. Shapiro,B. T. Darras. (2005). *Association of duchenne muscular dystrophy with autism spectrum disorder.* J Child Neurol, 20:790795.

[8] J.C . Denny, M. D. Ritchie, M. A. Basford,J.M. Pulley,L. Bastarache, K. Brown-Gentry, D. Wang,D. Masys, D. R. Roden,D. C. Crawford. (2010). *Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations..* Bioinformatics, 26(9):123-133.

Appendix A
Hierarchical Clustering

Hierarchical clustering (also known as connectivity base clustering) is a cluster analysis method that groups objects based off of how close together they are. The MATLAB hierarchical clustering method uses agglomerative clustering, which means that it starts by pairing similar patients to each other to form a cluster, and then pairs those clusters together until all patients are in one group.

We used Euclidean distance to measure similarity between patients. The linkage criteria determines which clusters to merge together. We used Ward's criterion, which looks at decreasing the variance for merged clusters. Because agglomerative clustering is $O(n^3)$, we tried to decrease the number of comparisons made by first running k-NN clustering for $k = 10$. Using k-NN means that a patient only needs to be compared to ten other patients instead of all the patients.

We considered other clustering methods including spectral clustering, which uses eigenvalues to group patients. We decided to use hierarchical clustering because it is relatively simple compared to other methods. Future work could include trying another clustering method.

Appendix B
Ignore CUIs

Table 2 shows the complete list of 34 CUIs that were ignored when calculating k-NN. As we can see, these CUIs tend to be very broad and were removed to try to show underlying relationships between patients.

TABLE II
Table of Ignored CUIs

CUI	Description	CUI	Description
C1533734	Administration proc.	C0001554	Administration act.
C0002778	Analysis of sub.	C0004339	Auscultation
C0005615	Birth	C0007634	Cells
C0011900	Diagnosis	C0012634	Disease
C0013658	Educational status	C0022885	Laboratory proc.
C0037088	Signs & symptoms	C0039082	Syndrome
C0043227	Work	C0087111	Therapeutic proc.
C0184661	Interventional	C0241889	Family history
C0243095	Finding	C0262926	Medical history
C0277786	Chief complaint	C0311392	Physical findings
C0332124	No past history	C0421451	Patient DoB
C0449416	Past medical history	C0455458	Pressure
C0518766	Vital signs	C0557061	Discussion
C0557854	Services	C0557985	Observation regimes
C0587081	Lab test finding	C0589120	Follow-up status
C1273870	Management proc.	C1299586	Difficulty
C1301826	Street address	C1457887	Symptoms

Appendix C
TF-IDF

Term frequency-inverse document frequency (tf-idf) is a statistical method that weights terms based off of how often they appear in a group of documents. Specifically, a term that appears often in a document but does not occur often among all the documents will be weighted higher.

The tf-idf value consists of term frequency multiplied by inverse document frequency.

The first component, term frequency tf(t, d), is the number of times that a term t appears in document d. For our calculations, tf(t, d) is the number of times a patient had a specific CUI.

The second component, inverse document frequency idf(t, D), assesses how often a term appears within all the documents. The specific equation: idf(t, D) = log $\frac{N}{|\{d \in D : t \in d\}|}$ In our calculations, N, the total number of patients, is 3625.

Finally, the tf-idf value is obtained by the following equation: tf_idf(t, d, D) = tf(t, d) x idf(t, D)