Data Science Seminars: Week 1 project

DISNET: Data mining project

Anton Aba: anton.abav@alumnos.upm.es
Ostap Kharysh: ostap.kharysh@alumnos.upm.es

Department of Computer Science
Master EIT Digital in Data Science

February 28, 2021

Contents

1 Scope

In this report we present our research on disease similarity analysis. With our approach we want to understand whether:

- the data we want to apply for our analysis is suitable for tasks of similarity analysis
- the approaches of similarity analysis provide valuable results

The data we want to access for the research is obtained from DISNET which is a service that acquires and updates the phenotypic manifestations of diseases by extracting them mainly from Wikipedia.

Here we will explain how the we collect the data, preprocess it and the reason for selecting algorithms for the similarity analysis tasks. As a result of our work we will prove the efficiency of using disease information gathered by DISNET and how efficient it could be for relational analysis of the diseases.

2 Data preprocessing

To begin with, as we want to analyze the disease similarities we need to acquire as many as possible values to characterize each of them. For this reason we went through DISNET service list and decided to use this query "*diseaseWithMoreDisnetConcepts*" which provides diseases with the highest possible amount of information, that includes phenotypic manifestations. Also, we decided that 200 diseases is relatively enough for purposes of our research.

As one could infer from the 1, after the series of programming manipulations, we acquired a table with 200 rows as number of diseases and their 381 genotypic manifestations, each identified as 1 if exists and 0 otherwise.

	name	Loose stool	Pruritus ani	Numbness	Rhinorrhea	Pruritus	Pain in testicle	Joint swelling	Bulky stool	Ataxia	Multiple sleep latency test	Extrapyramidal rigidity
0	Kawasaki disease	0	0	0	1	1	0	0	0	0	0	0
1	Cerebral palsy	0	0	0	0	0	0	0	0	0	0	0
2	Heart failure	0	0	0	0	0	0	0	0	0	0	0
3	Hypoglycemia	0	0	1	0	0	0	0	0	1	0	0
4	Anorexia nervosa	0	0	0	0	0	0	0	0	0	0	0

Figure 1: Table containing Binary hot-encoding of diseases

3 Clustering Algorithms

After careful consideration of possible similarity analysis approaches we decided not to conduct relation analysis based on semantic similarities of the names of diseases or their features. Thus, we took a closer look at clustering techniques which we believe will be effectively exploit the possible 381 genotypic manifestations applying them.

For our particular case we decided to implement two clustering algorithms and compare their outputs, in order to find those sets of diseases which were highly connected: the K-Means and the Hierarchical Agglomerative Clustering algorithms.

Our believe is that if 2 clusters corresponding to different algorithms share a big amount of diseases, it would evidence the connection between those shared affections based on the features gathered on DISNET. In order to properly explain our findings we will discuss and annex the results of an example execution.

3.1 K-Means Algorithm

This algorithm is requires to predefine a number of clusters, allocates the centroids randomly and does simultaneous grouping based on the distance in the hyperplane-based of the points. It is either they belong to a cluster or not, with no defined similarities of hierarchy with the other classes, separate entities.

As we do not hold related medical expertise it is difficult to determine the number of possible disease taxonomies. This number is determined by the diseases chosen to be clustered, as well as the initial selection of centroids (which is random). Therefore, it may vary on each particular execution adapting to each respective input.

To estimate the number of clusters to use we apply the **Elbow method**, which consists of comparing the standard square error of models built with different number of clusters. The testing range is from 1 to 15 centroids, since it did not seem reasonable to make an exaggerated amount of clusters which may lead to overfitting.

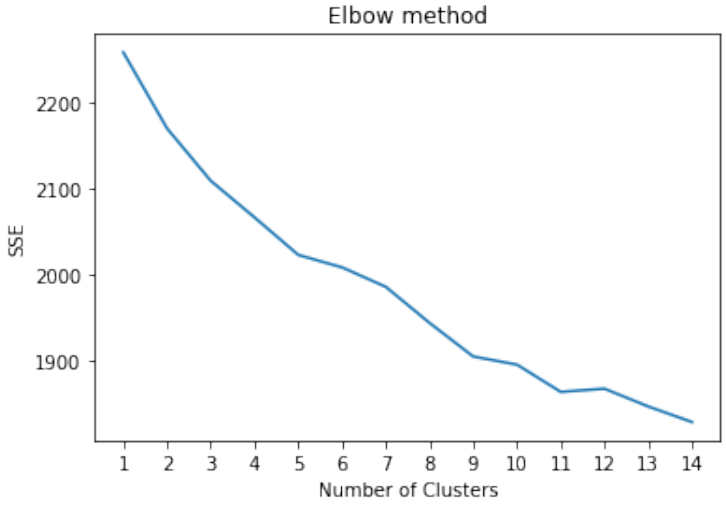

Figure 2: Elbow method visually

The desired number of clusters seems to be around 6, depending on the execution details. Based on obtained clusters one could claim that each of them is related to the specific group of diseases but with an uncertain amount of false positives.

As an example, the next image shows the results of this algorithm simplified to a 2-Dimensional space. It is clear that, although there are a few outliers, diseases which are close to each other in the hyperplane are grouped together by the algorithm.

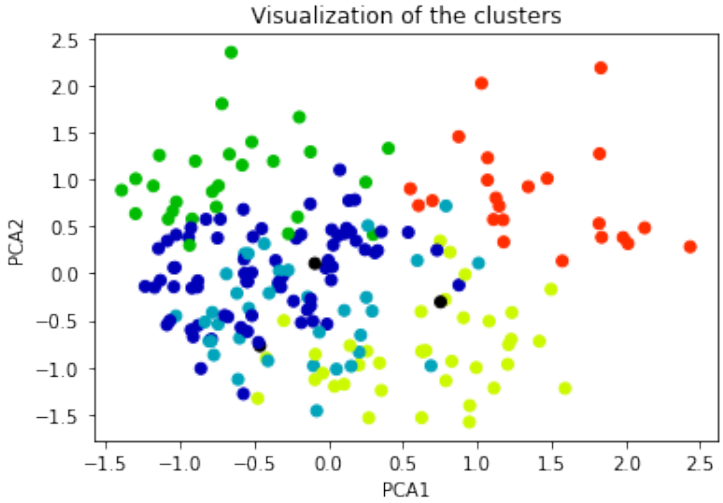

Figure 3: PCA for K-Means

3.2 Hierarchical Agglomerative Clustering

We picked this algorithm because it provides a sequential approach to hierarchical clustering. This reminded us of the biological classification approach, in which all living beings belong to a family, class, subclass, etc. It seemed reasonable that disease classification could be done in a similar fashion, based on their phenotypical expressions.

Looking at the Dendrogram and at the results from the elbow method the number of clusters is determined to be the same as with K-Means. It makes sense that the number of clusters is the same or close to the same, otherwise one of the methods must be implemented wrong. As it can be seen below, the number of clusters makes sense, making it bigger would be to specific and smaller too general and the diseases are grouped, subgrouped and related in different degrees of similarity, as they are in real life.

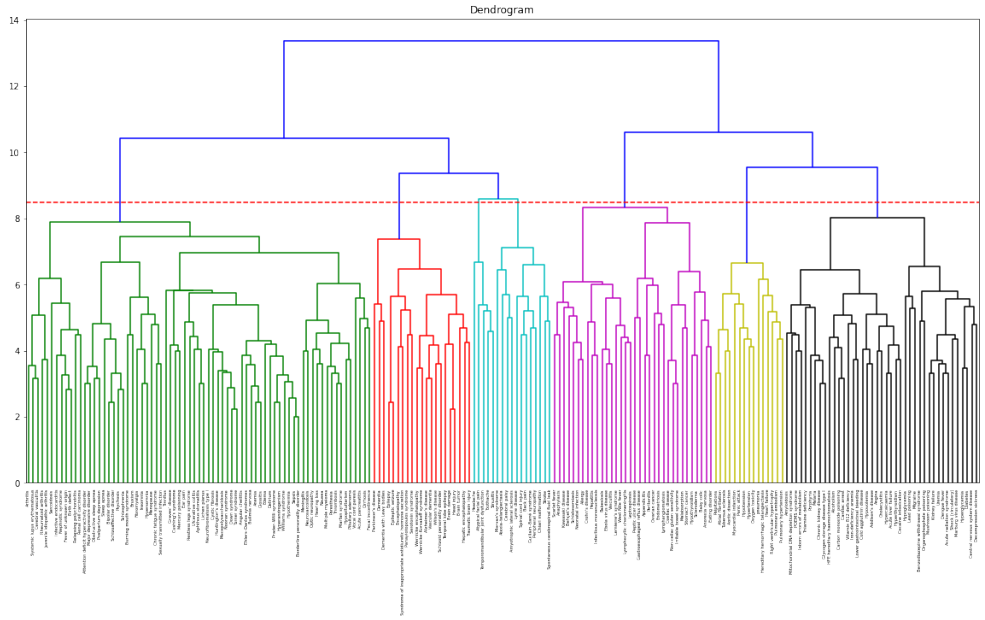

Figure 4: Dendrogram of the Hierarchical Clustering

Below, on the other hand we can see how the 2D reduction of the space shows that clusters are correctly computed. With a similar certainty of its correctness to the K-Means example, therefore it will be interesting to see which clusters are very similar and which are not.

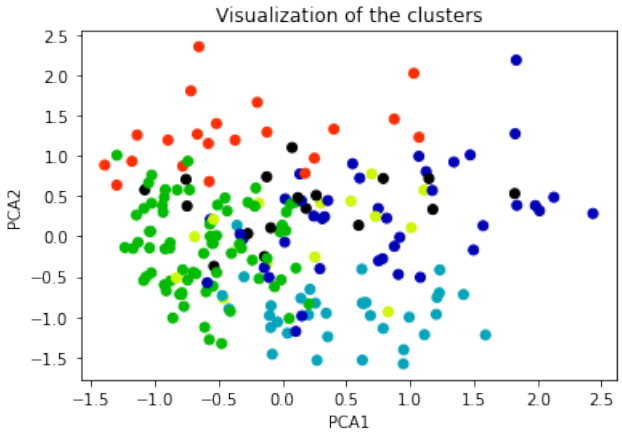

Figure 5: PCA visualization of Hierarchical clustering

4 Analysis of results

4.1 Comparison

Using the Jaccard coefficient between clusters and keeping those over 50% in order to find subclusters of very similiar diseases. We will keep those subclusters in order to prove the validity of our methodology.

If $J(A,B) = \frac{|A \cap B|}{|A \cup B|} >= 0.5$ then there is high similarity between clusters of distinct algorithms and therefore it is determined that the items resulting from the intersection most definitely are members of the same family of diseases, phenotypically-wise.

4.2 KMeans - Hierarchical Agglomeration similarities

KMeans Cluster 1 elements are 64.86% of the times in Agglomeration Cluster 3, which represent 65.75% of Cluster 3's size. That is a strong indicator that these diseases belonging to both groups must be heavily phenotipically related. Similarity Clusters 1 and 3 : 65.31%. It clusters a set of Cognitively affected (retardation, personality, sleeping) and general inflammatory related diseases.

KMeans Cluster 3 out of all its items, 59.26% of them are present in the Agglomeration Cluster 5, which 76.19% of it are those diseases. Clusters 3 and 5 : 66.67%. It clusters a set of Cognitive deterioration related affections, such as different kinds of dementia or epilepsy

Furthermore, KMeans Cluster 4 elements are 73.68% of the time in Cluster 2, which are 82.35% of its own total. Clusters 4 and 2 : 77.78%. Diseases related to gastrointestinal irritation and infection.

5 Conclusions

Based on the *Similarity analysis* section we concluded that the data obtained from DISNET could be used for disease clustering as clustering algorithms were able to identify several meaningful clusters of certain types of diseases.

Moreover, based on comparison of the clusters produced by distinct algorithms one could infer that some of them share a significant amount of the same algorithms which advises that the results obtained are reasonably related to a specific topic and are not combined by chance.

In conclusion, we presume that DISNET is well-justified information resource for implementing similarity analysis of diseases, for instance, resulting in meaningful results based on clustering approaches described in this research.

6 Annex

6.1 Results of K-Means

Cluster 0 (size = 3): 'Thiamine deficiency', 'Guillain–Barré syndrome', 'Right ventricular hypertrophy'.

Cluster 1 (size = 74): 'Cerebral palsy', 'Nephrotic syndrome', 'Stroke', 'Attention deficit hyperactivity disorder', 'Anemia', 'Palpitations', 'Delirium', 'Meningitis', 'Hypothyroidism', 'Fever of unknown origin', 'Fecal incontinence', 'Major depressive disorder', 'Iron-deficiency anemia', 'Angular cheilitis', 'Spinal cord injury', 'Ulcerative colitis', 'Insomnia', 'Obstructive sleep apnea', 'Chronic kidney disease', 'Optic neuropathy', 'Ear pain', "Graves' disease", 'Hypopituitarism', 'Hypersomnia', 'Sepsis', 'Thrombosis', 'Lower gastrointestinal bleeding', 'Edema', 'Down syndrome', "Cushing's syndrome", 'Hearing loss', 'POEMS syndrome', 'Mucopolysaccharidosis', 'Burning mouth syndrome', 'Bulimia nervosa', 'Porphyria', 'Hypercalcaemia', 'Birth defect', 'Cardiac arrest', 'Cystic fibrosis', 'Acute liver failure', 'Tuberous sclerosis', 'Marfan syndrome', 'Borderline personality disorder', 'Cocaine intoxication', 'Paresthesia', 'Mitochondrial DNA depletion syndrome', 'Glossitis', 'Schizoid personality disorder', 'Mercury poisoning', 'Malaria', 'Nausea', 'Menopause', 'Amyotrophic lateral sclerosis', 'Neurofibromatosis type I', 'Hypothermia', 'Amyloidosis', 'Noonan syndrome', 'Vitamin B12 deficiency', 'Oxygen toxicity', 'Williams syndrome', 'Lichen planus', 'Turner syndrome', 'Postpartum depression', 'Multiple myeloma', 'Sexually transmitted infection', 'Tinnitus', 'Syndrome of inappropriate antidiuretic hormone secretion', 'Prader–Willi syndrome', 'Restless legs syndrome', 'Hypotension', 'Vasculitis', 'Hyperandrogenism', 'Peripheral neuropathy'.

Cluster 2 (size = 36): 'Heart failure', 'Cirrhosis', 'Sarcoidosis', 'Hepatitis', 'Acute pancreatitis', 'Systemic lupus erythematosus', 'Rheumatoid arthritis', 'Relapsing polychondritis', 'Ovarian cancer', 'Myocardial infarction', 'Atrial fibrillation', 'Chiari malformation', 'Ehlers–Danlos syndromes', 'HFE hereditary haemochromatosis', 'Toothache', 'Temporomandibular joint dysfunction', 'Atypical facial pain', 'Appendicitis', 'Back pain', 'Glycogen storage disease type I', 'Juvenile idiopathic arthritis', 'Bruxism', 'Pulmonary embolism', 'Pituitary adenoma', 'Chronic fatigue syndrome', 'Infective endocarditis', 'Sinusitis', 'Vocal cord paresis', 'Pulmonary hypertension', 'Arthritis', 'Angina', 'Renal cell carcinoma', 'Aortic dissection', 'Babesiosis', 'Cerebral vasculitis', 'Cholecystitis'.

Cluster 3 (size = 27): 'Dementia', 'Psychosis', 'Dementia with Lewy bodies', 'Hepatic encephalopathy', 'Wernicke encephalopathy', 'Ataxia–telangiectasia', "Parkinson's disease", 'Panayiotopoulos syndrome', 'Hereditary hemorrhagic telangiectasia', 'Epilepsy', "Wilson's disease", 'Schizophrenia', 'Epileptic seizure', 'Sleep apnea', "Morvan's syndrome", 'Bipolar disorder', "Alzheimer's disease", 'Encephalopathy', 'Temporal lobe epilepsy', 'Wernicke–Korsakoff syndrome', "Huntington's disease", 'Fibromyalgia', 'Brain tumor', 'Vascular dementia', 'Schizoaffective disorder', 'Neurosarcoidosis', 'Serotonin syndrome'.

Cluster 4 (size = 38): 'Kawasaki disease', 'Anorexia nervosa', "Crohn's disease", 'Eating disorder', 'Hyperthyroidism', "Behçet's disease", 'pneumonia', 'Shock (circulatory)', 'Lymphocytic choriomeningitis', 'Leptospirosis', 'Malabsorption', 'Peptic ulcer disease', 'Baby colic', 'Scarlet fever', 'Non-celiac gluten sensitivity', "Lemierre's syndrome", 'Marburg virus disease', 'Inborn errors of metabolism', 'Irritable bowel syndrome', 'Peritonitis', 'Neonatal infection', 'Schistosomiasis', 'Coeliac disease', 'Anaphylaxis', 'Dengue fever', 'Endometriosis', 'Scleroderma', 'West Nile fever', 'Aphthous stomatitis', 'Allergy', 'Cold agglutinin disease', "Addison's disease", 'Kidney failure', 'Ebola virus disease', 'In-

fectious mononucleosis', 'Lymphangiomatosis', 'Gastroesophageal reflux disease', 'Stomach cancer'.

Cluster 5 (size = 23): 'Hypoglycemia', 'Lead poisoning', 'Headache', 'Lyme disease', 'Hypertension', 'Uremia', 'Brain damage', 'Alcoholism', 'Migraine', 'Benzodiazepine withdrawal syndrome', 'Central nervous system disease', 'Hyperglycemia', 'Head injury', 'Carbon monoxide poisoning', 'Diabetes', 'Organophosphate poisoning', 'Panic attack', 'Mushroom poisoning', 'Hypovolemia', 'Acute radiation syndrome', 'Spontaneous cerebrospinal fluid leak', 'Traumatic brain injury', 'Decompression sickness'.

6.2 Agglomerative Clustering results

Cluster 0 (size = 17): 'Cerebral palsy', 'Headache', 'Lyme disease', 'Stroke', 'Ataxia–telangiectasia', 'Chiari malformation', 'Spinal cord injury', 'Toothache', 'Temporomandibular joint dysfunction', "Morvan's syndrome", 'Atypical facial pain', 'Back pain', 'Sinusitis', 'Guillain–Barré syndrome', 'Amyotrophic lateral sclerosis', 'Spontaneous cerebrospinal fluid leak', 'Peripheral neuropathy'.

Cluster 1 (size = 41): 'Hypoglycemia', 'Lead poisoning', 'Hyperthyroidism', 'Uremia', 'Alcoholism', 'Shock (circulatory)', 'Migraine', 'Iron-deficiency anemia', 'Benzodiazepine withdrawal syndrome', 'Central nervous system disease', 'Hyperglycemia', 'HFE hereditary haemochromatosis', 'Chronic kidney disease', 'Thiamine deficiency', 'Carbon monoxide poisoning', 'Diabetes', 'Lower gastrointestinal bleeding', 'POEMS syndrome', 'Glycogen storage disease type I', 'Marburg virus disease', 'Inborn errors of metabolism', 'Organophosphate poisoning', 'Porphyria', 'Hypercalcaemia', 'Cardiac arrest', 'Acute liver failure', 'Cocaine intoxication', 'Mitochondrial DNA depletion syndrome', 'Dengue fever', 'Malaria', 'Mushroom poisoning', 'Hypovolemia', 'Amyloidosis', 'Vitamin B12 deficiency', 'Angina', 'Cold agglutinin disease', "Addison's disease", 'Acute radiation syndrome', 'Kidney failure', 'Decompression sickness', 'Cholecystitis'.

Cluster 2 (size = 34): 'Kawasaki disease', 'Anorexia nervosa', "Crohn's disease", 'Hepatitis', 'Eating disorder', "Behçet's disease", 'Hypothyroidism', 'Lymphocytic choriomeningitis', 'Ovarian cancer', 'Leptospirosis', 'Malabsorption', 'Peptic ulcer disease', 'Baby colic', 'Appendicitis', 'Scarlet fever', 'Non-celiac gluten sensitivity', "Lemierre's syndrome", 'Irritable bowel syndrome', 'Peritonitis', 'Neonatal infection', 'Schistosomiasis', 'Coeliac disease', 'Anaphylaxis', 'Endometriosis', 'Scleroderma', 'West Nile fever', 'Allergy', 'Babesiosis', 'Ebola virus disease', 'Vasculitis', 'Infectious mononucleosis', 'Lymphangiomatosis', 'Gastroesophageal reflux disease', 'Stomach cancer'.

Cluster 3 (size = 73): 'Cirrhosis', 'Sarcoidosis', 'Nephrotic syndrome', 'Acute pancreatitis', 'Psychosis', 'Systemic lupus erythematosus', 'Attention deficit hyperactivity disorder', 'Anemia', 'Rheumatoid arthritis', 'Delirium', 'Meningitis', 'Relapsing polychondritis', 'Fever of unknown origin', 'Fecal incontinence', 'Major depressive disorder', 'Angular cheilitis', 'Ehlers–Danlos syndromes', 'Ulcerative colitis', 'Insomnia', 'Obstructive sleep apnea', 'Schizophrenia', 'Optic neuropathy', 'Ear pain', "Graves' disease", 'Hypopituitarism', 'Sleep apnea', 'Hypersomnia', 'Sepsis', 'Thrombosis', 'Edema', 'Down syndrome', "Cushing's syndrome", 'Hearing loss', 'Mucopolysaccharidosis', 'Burning mouth

syndrome', 'Juvenile idiopathic arthritis', 'Bruxism', 'Bulimia nervosa', 'Bipolar disorder', 'Birth defect', 'Cystic fibrosis', 'Pituitary adenoma', 'Chronic fatigue syndrome', 'Marfan syndrome', 'Borderline personality disorder', 'Infective endocarditis', 'Paresthesia', 'Glossitis', 'Vocal cord paresis', 'Mercury poisoning', 'Arthritis', 'Nausea', "Huntington's disease", 'Menopause', 'Fibromyalgia', 'Aphthous stomatitis', 'Neurofibromatosis type I', 'Hypothermia', 'Noonan syndrome', 'Williams syndrome', 'Renal cell carcinoma', 'Lichen planus', 'Turner syndrome', 'Postpartum depression', 'Multiple myeloma', 'Sexually transmitted infection', 'Schizoaffective disorder', 'Cerebral vasculitis', 'Tinnitus', 'Prader–Willi syndrome', 'Restless legs syndrome', 'Neurosarcoidosis', 'Hyperandrogenism'.

Cluster 4 (size = 15): 'Heart failure', 'Hypertension', 'pneumonia', 'Palpitations', 'Hereditary hemorrhagic telangiectasia', 'Myocardial infarction', 'Atrial fibrillation', 'Pulmonary embolism', 'Tuberous sclerosis', 'Pulmonary hypertension', 'Panic attack', 'Oxygen toxicity', 'Aortic dissection', 'Right ventricular hypertrophy', 'Hypotension'.

Cluster 5 (size = 21): 'Dementia', 'Dementia with Lewy bodies', 'Hepatic encephalopathy', 'Brain damage', 'Wernicke encephalopathy', "Parkinson's disease", 'Panayiotopoulos syndrome', 'Epilepsy', 'Head injury', "Wilson's disease", 'Epileptic seizure', "Alzheimer's disease", 'Encephalopathy', 'Schizoid personality disorder', 'Temporal lobe epilepsy', 'Wernicke–Korsakoff syndrome', 'Brain tumor', 'Vascular dementia', 'Traumatic brain injury', 'Syndrome of inappropriate antidiuretic hormone secretion', 'Serotonin syndrome'.

6.3 Similar clusters

K-Means Cluster 1 and Agglomerative Custer 3 (size = 48): 'Nephrotic syndrome', 'Attention deficit hyperactivity disorder', 'Anemia', 'Delirium', 'Meningitis', 'Fever of unknown origin', 'Fecal incontinence', 'Major depressive disorder', 'Angular cheilitis', 'Ulcerative colitis', 'Insomnia', 'Obstructive sleep apnea', 'Optic neuropathy', 'Ear pain', "Graves' disease", 'Hypopituitarism', 'Hypersomnia', 'Sepsis', 'Thrombosis', 'Edema', 'Down syndrome', "Cushing's syndrome", 'Hearing loss', 'Mucopolysaccharidosis', 'Burning mouth syndrome', 'Bulimia nervosa', 'Birth defect', 'Cystic fibrosis', 'Marfan syndrome', 'Borderline personality disorder', 'Paresthesia', 'Glossitis', 'Mercury poisoning', 'Nausea', 'Menopause', 'Neurofibromatosis type I', 'Hypothermia', 'Noonan syndrome', 'Williams syndrome', 'Lichen planus', 'Turner syndrome', 'Postpartum depression', 'Multiple myeloma', 'Sexually transmitted infection', 'Tinnitus', 'Prader–Willi syndrome', 'Restless legs syndrome', 'Hyperandrogenism'. Sleeping / Attention / Personality / (Borned) Intellectual disability -

K-Means Cluster 3 and Agglomerative Custer 5 (size = 16):'Dementia', 'Dementia with Lewy bodies', 'Hepatic encephalopathy', 'Wernicke encephalopathy', "Parkinson's disease", 'Panayiotopoulos syndrome', 'Epilepsy', "Wilson's disease", 'Epileptic seizure', "Alzheimer's disease", 'Encephalopathy', 'Temporal lobe epilepsy', 'Wernicke–Korsakoff syndrome', 'Brain tumor', 'Vascular dementia', 'Serotonin syndrome'.

K-Means Cluster 4 and Agglomerative Custer 2 (size = 28): 'Kawasaki disease', 'Anorexia nervosa', "Crohn's disease", 'Eating disorder', "Behçet's disease", 'Lym-

phocytic choriomeningitis', 'Leptospirosis', 'Malabsorption', 'Peptic ulcer disease', 'Baby colic', 'Scarlet fever', 'Non-celiac gluten sensitivity', "Lemierre's syndrome", 'Irritable bowel syndrome', 'Peritonitis', 'Neonatal infection', 'Schistosomiasis', 'Coeliac disease', 'Anaphylaxis', 'Endometriosis', 'Scleroderma', 'West Nile fever', 'Allergy', 'Ebola virus disease', 'Infectious mononucleosis', 'Lymphangiomatosis', 'Gastroesophageal reflux disease', 'Stomach cancer'. Gastrointestinal