Weighted Symptom Disease network

Team members:

Akash Gangadharan (akganga@iu.edu)

Sandhya Nagarajan Iyer (sniyer@iu.edu)

Github link: https://github.com/SandhyaNIyer/Weighted-Symptom-Disease-network

Abstract:

Symptom disease networks provide an invaluable tool for identifying disease associations and predicting patient outcomes, as they help to uncover the complex web of connections that exist between different diseases and their symptoms. In this study, we utilized the comprehensive DoPathways (https://snap.stanford.edu/biodata/datasets/10005/10005-D-DoPathways.html) dataset to construct a weighted symptom disease network, with each edge representing a symptom shared by two diseases and the weight of each edge reflecting the number of shared symptoms. The Weighted Symptoms-Disease Network will help us understand the complex relationships between diseases and their associated symptoms.

By employing network analysis methods such as degree, centrality measures, clustering coefficient, and community detection, we were able to explore the structure of the network and identify key nodes that play critical roles in connecting different parts of the network. Our primary goal was to compare the performance of the inbuilt community detection algorithm in Gephi, a widely used network analysis software, with the actual disease categories.

Our results indicate that while the inbuilt community detection algorithm was able to create disease communities that generally corresponded to the actual disease categories, there were some discrepancies. Our study also identified diseases that were not assigned to their correct category, indicating potential areas for improvement in the community detection algorithm.

Introduction:

The identification of disease associations and the prediction of patient outcomes are important goals in healthcare. Symptom disease networks are a promising approach to achieving these goals, as they allow us to explore the relationships between diseases and their associated symptoms. In this study, we constructed a weighted symptom disease network using the DoPathways dataset, which only contains various diseases and their categories. It includes categories for each disease, such as Acquired Metabolic disease, Benign Neoplasm, Cancer, Cardiovascular System disease, Gastrointestinal System disease, Immune System disease, Inherited Metabolic disorder, Musculoskeletal System disease, Nervous System Disease, Respiratory System Disease which

have been manually collected from the Center for Disease Control and other government websites and used to construct the network. In this network, diseases are represented as nodes, and symptoms are represented as edges. The weight of each edge corresponds to the number of symptoms shared by the two diseases.

Motivation:

Our main motivation for this research is to evaluate the effectiveness of the community detection method of Gephi in correctly classifying diseases into their respective categories. Specifically, we aim to identify any diseases that are incorrectly classified and explore the factors that may contribute to this misclassification.

We used network analysis techniques, including degree distribution, centrality measures, and community detection, to explore the structure of the symptom disease network and identify key diseases and how they relate to other diseases in the network. Our analysis focused on evaluating the accuracy of the inbuilt community detection algorithm in Gephi software in correctly assigning each disease to its respective category.

The results of our study have important implications for disease diagnosis and management. Identifying disease associations and predicting patient outcomes can inform precision medicine approaches that are tailored to individual patients based on their symptom profiles. Additionally, understanding the factors that contribute to misclassification of diseases can help improve the accuracy of disease classification algorithms, leading to more effective disease management and treatment.

Objectives:

- Construct a weighted symptom disease network using the DoPathways dataset.
- Perform network analysis techniques, including degree distribution, centrality measures, and community detection, to explore the structure of the network and identify key diseases and how they relate to other diseases in the network.
- Identify diseases that are incorrectly classified and explore the factors that may contribute to this misclassification.

Hypotheses:

- The weighted symptom disease network will help us understand the complex relationships between diseases and their associated symptoms.
- The community detection algorithm will be able to accurately classify diseases into their respective categories.
- There may be diseases that are misclassified due to shared symptoms or other factors.

Methods:

The methodology or the approach to this project is done in two phases. The first phase is Exploratory Data analysis, and the second phase consists of Network Analysis and community detection in Gephi.

Phase 1:

Exploratory Data Analysis:

The Exploratory Data analysis consists of 4 stages.

1) Data Collection
2) Data preprocessing
3) Visualization
4) Insights

Data Collection:

We collected the data from the (DoPathways). This data set consists of Diseases and its Categories. The categories are Acquired Metabolic disease, Benign Neoplasm, Cancer, Cardiovascular System disease, Gastrointestinal System disease, Immune System disease, Inherited Metabolic disorder, Musculoskeletal System disease, Nervous System Disease, Respiratory System Disease.

Data Preprocessing:

The data we collected from the DoPathways only contains Diseases and its categories. However, we also wanted symptoms of those diseases since the whole project is based on the diseases and their symptoms. We manually collected symptoms for each disease from the Centre of disease control and other government websites to make sure we have the right data. Then we merged these symptoms data with our original data that has diseases and its categories.

Data Visualization:

Once we got the data preprocessed, we performed Exploratory Data analysis to get insights from the data. Data visualization also helped us to validate the findings and the study from other researcher papers and websites.

Insights:

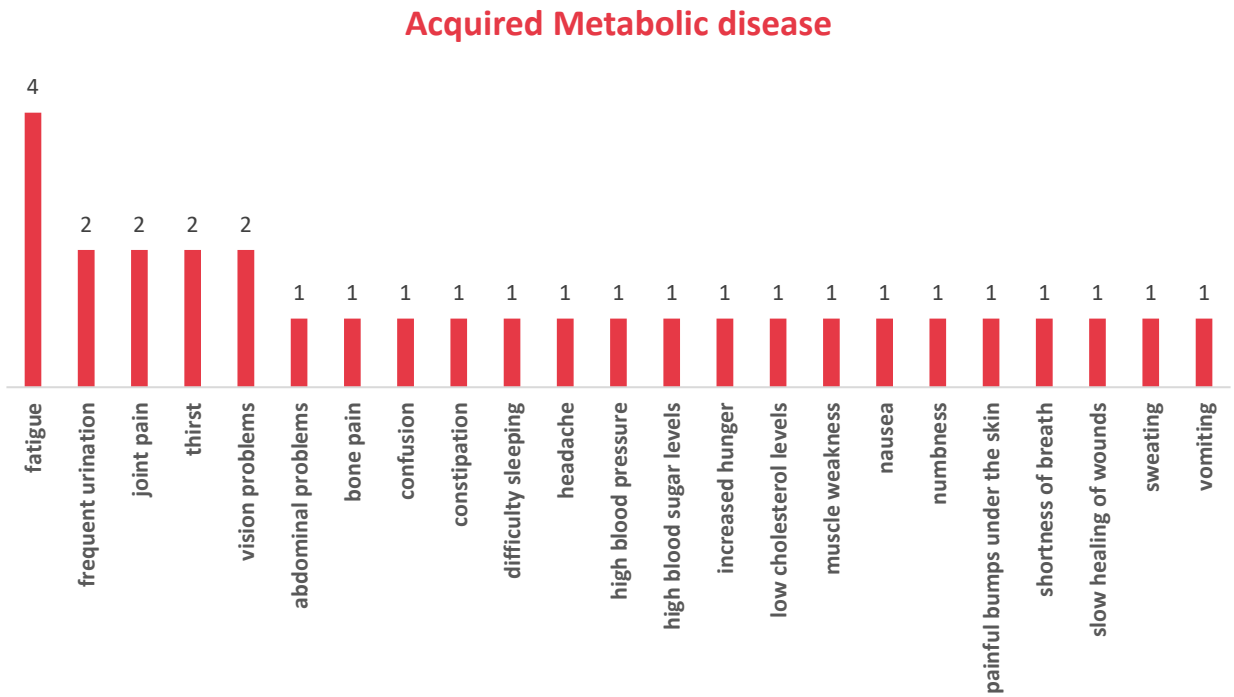
From the data visualization, we then generated some important insights. The graphs below show the most prevalent symptoms in the diseases of a category. Furthermore, information on symptom frequency can help healthcare providers in their diagnostic process. By identifying the most common symptoms, healthcare providers may be able to narrow down potential diagnoses and provide more effective treatments for that disease category.

Insights from Exploratory Data Analysis:

1. **Acquired Metabolic disease:**

In the below chart, we can see that fatigue, frequent urination, and joint pain are common symptoms across various acquired metabolic disorders, including diabetes mellitus, metabolic syndrome X, and obesity.

Fatigue can be related to the body's inability to effectively use glucose for energy, which can occur in diabetes mellitus and metabolic syndrome X. In obesity, excess weight can lead to physical strain on the body, which can cause fatigue.



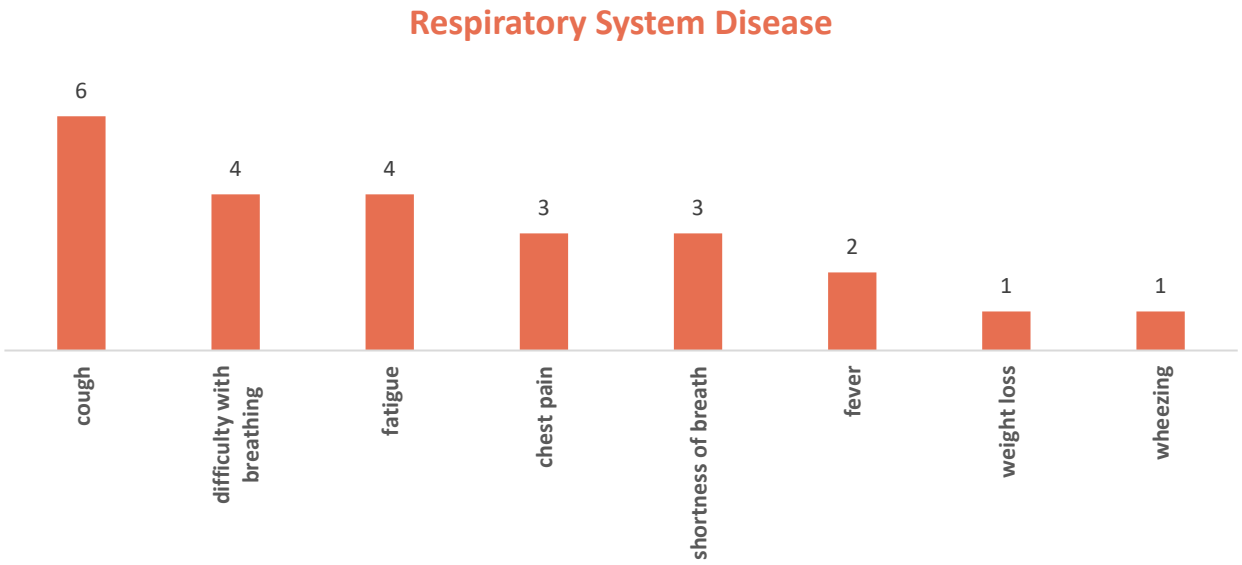
Frequent urination can occur in diabetes mellitus and hyperglycemia due to the increased production of urine as the body tries to rid itself of excess glucose. In obesity, increased pressure on the bladder due to excess weight can cause frequent urination.

Joint pain can be related to the inflammation that can occur in metabolic syndrome X and obesity. Obesity can also lead to mechanical stress on joints, causing pain and inflammation. In diabetes mellitus, joint pain can occur due to nerve damage and poor circulation.

2. **Respiratory System Disease:**

Cough, difficulty in breathing, and fatigue are common symptoms in various respiratory diseases, including asthma, chronic obstructive pulmonary disease (COPD), interstitial lung diseases, pneumonia, and pulmonary fibrosis.

Cough is a reflex action that helps to clear mucus, irritants, or foreign particles from the respiratory tract. In respiratory diseases, cough can be caused by inflammation, infection, or obstruction in the airways, as well as the body's attempt to clear excess mucus or irritants from the lungs. Cough can be severe and persistent in some respiratory diseases, leading to fatigue and impaired quality of life. [11]



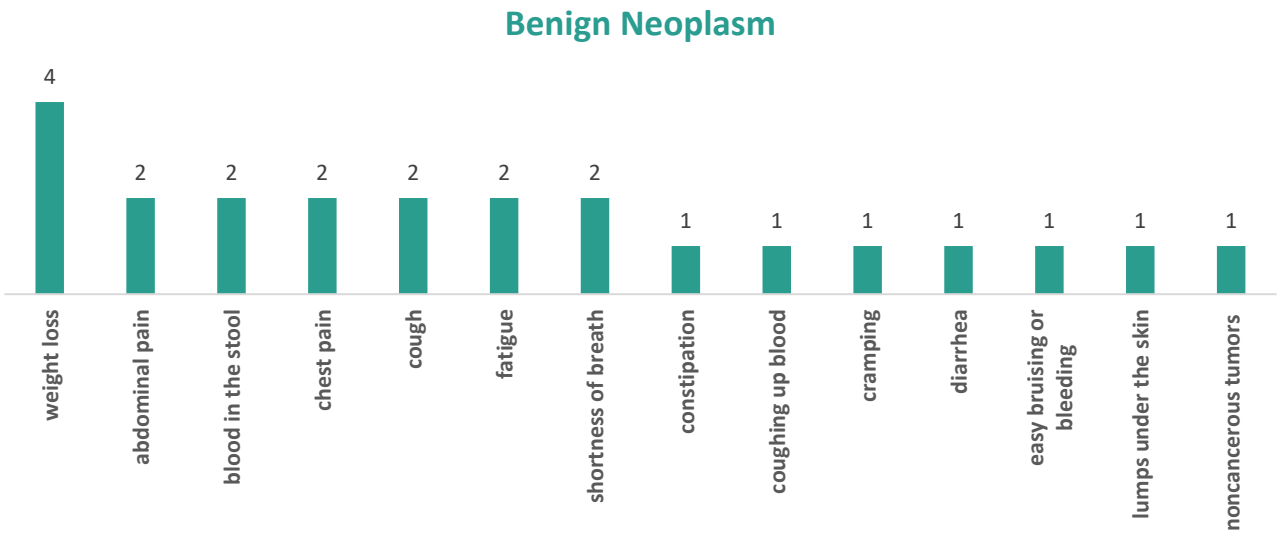
Difficulty in breathing, also known as dyspnea, can occur due to obstruction or narrowing of the airways, inflammation or scarring of lung tissue, or a decrease in lung function. In asthma, difficulty in breathing can be caused by inflammation and narrowing of the airways, while in COPD, it can be caused by damage to the lung tissue and narrowing of the airways. In interstitial lung diseases and pulmonary fibrosis, difficulty in breathing can be caused by scarring of the lung tissue, which can lead to a decrease in lung function and reduced oxygen uptake. [12]

Fatigue can occur in respiratory diseases due to the body's increased effort to breathe, as well as the impact of chronic inflammation on the body. In some respiratory diseases, such as interstitial

lung diseases and pulmonary fibrosis, fatigue can also be caused by reduced oxygen uptake, leading to decreased energy levels and impaired daily functioning.

3. Benign Neoplasm:

In the below chart, we can see that weight loss, abdominal pain, and blood in the stool occur the highest number of times in a Benign neoplasm symptom category, it suggests that these symptoms may be common in benign tumors that grow in the gastrointestinal tract. Benign neoplasms are abnormal growths of cells that are not cancerous, and they can develop in various parts of the body.



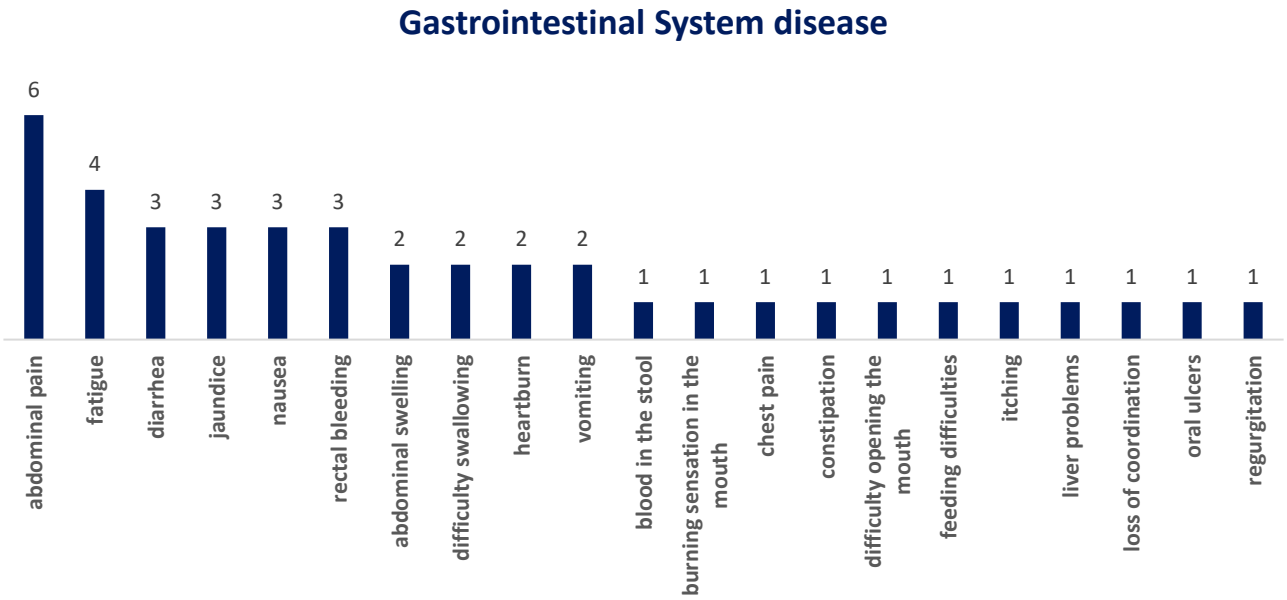
Weight loss can occur as a result of changes in metabolism due to the tumor or due to changes in appetite and digestion. Abdominal pain can occur due to the tumor compressing nearby organs or due to inflammation caused by the tumor. Blood in the stool can occur due to bleeding from the tumor or due to irritation of the gastrointestinal tract caused by the tumor.

4. Gastrointestinal System disease:

Abdominal pain, fatigue, and diarrhea are common symptoms in the Gastrointestinal System disease category because many of the diseases in this category affect the function of the gastrointestinal tract. The gastrointestinal tract is responsible for digesting and absorbing nutrients from food, and any disruption to its function can lead to a range of symptoms, including abdominal pain, fatigue, and diarrhea.

Abdominal pain can occur due to inflammation, irritation, or obstruction in the gastrointestinal tract. Inflammatory bowel diseases (such as colitis and ulcerative colitis) and necrotizing

enterocolitis are examples of conditions that can cause inflammation in the gastrointestinal tract, leading to abdominal pain.



Fatigue can occur as a result of changes in metabolism due to the disease or due to changes in sleep patterns and activity levels. For example, liver cirrhosis, hepatitis, and hepatolenticular degeneration are examples of conditions that can cause changes in metabolism due to damage to the liver, leading to fatigue.

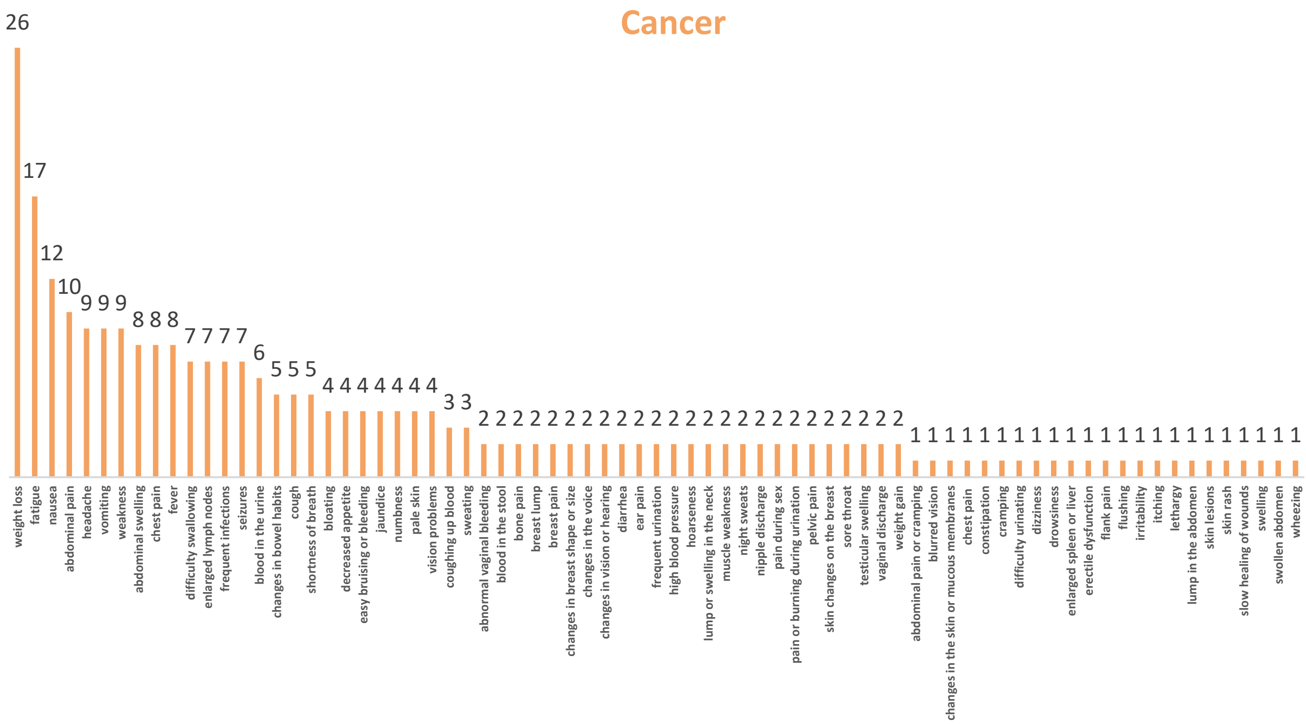
Diarrhea can occur due to inflammation in the gastrointestinal tract or due to changes in digestion and absorption of nutrients. Inflammatory bowel diseases and cholestasis are examples of conditions that can cause inflammation in the gastrointestinal tract, leading to diarrhea.

5. **Cancer:**

Weight loss, fatigue, and nausea are common symptoms in cancer patients due to several reasons. First, cancer cells consume a lot of energy and nutrients from the body, leading to weight loss and fatigue. Second, cancer treatment such as chemotherapy and radiation therapy can cause side effects like nausea, vomiting, and loss of appetite, leading to weight loss and fatigue. Third, cancer can cause inflammation in the body, leading to fatigue and other symptoms.

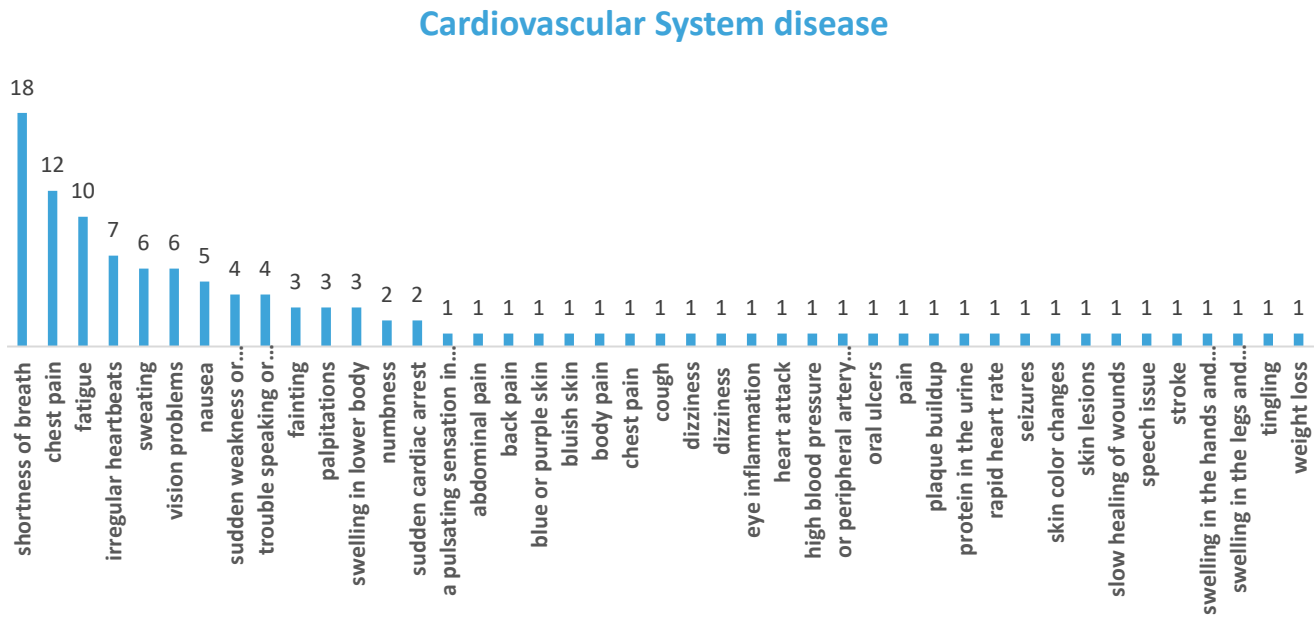
In addition to these general reasons, specific cancers may cause weight loss, fatigue, and nausea through various mechanisms. For eg., in pancreatic cancer, the tumor can cause obstruction of the digestive system, leading to weight loss and nausea. In some types of leukemia, cancer cells can

invade the bone marrow and disrupt the production of blood cells, leading to fatigue and weakness.



6. **Cardiovascular System Disease:**

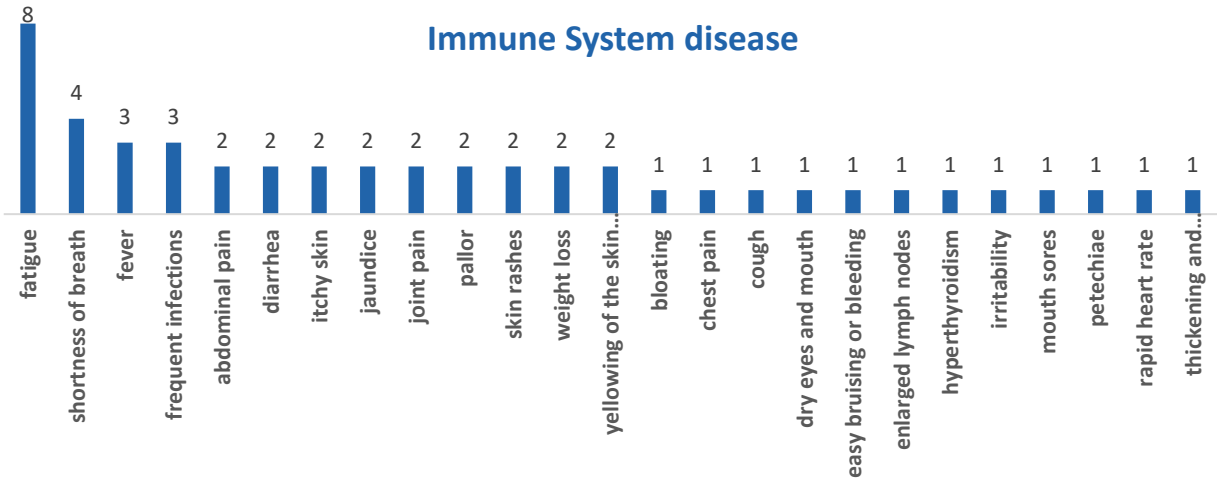
These symptoms can be indicative of the heart not receiving enough oxygen-rich blood, which can be caused by various underlying diseases. One insight you can gain from this observation is that these symptoms may not be specific to any cardiovascular disease, and further testing may be necessary to identify the underlying cause. For example, shortness of breath and chest pain can be symptoms of coronary artery disease or pulmonary hypertension, while irregular heartbeats can be a symptom of atrial fibrillation or a congenital heart defect.



7. Immune System disease:

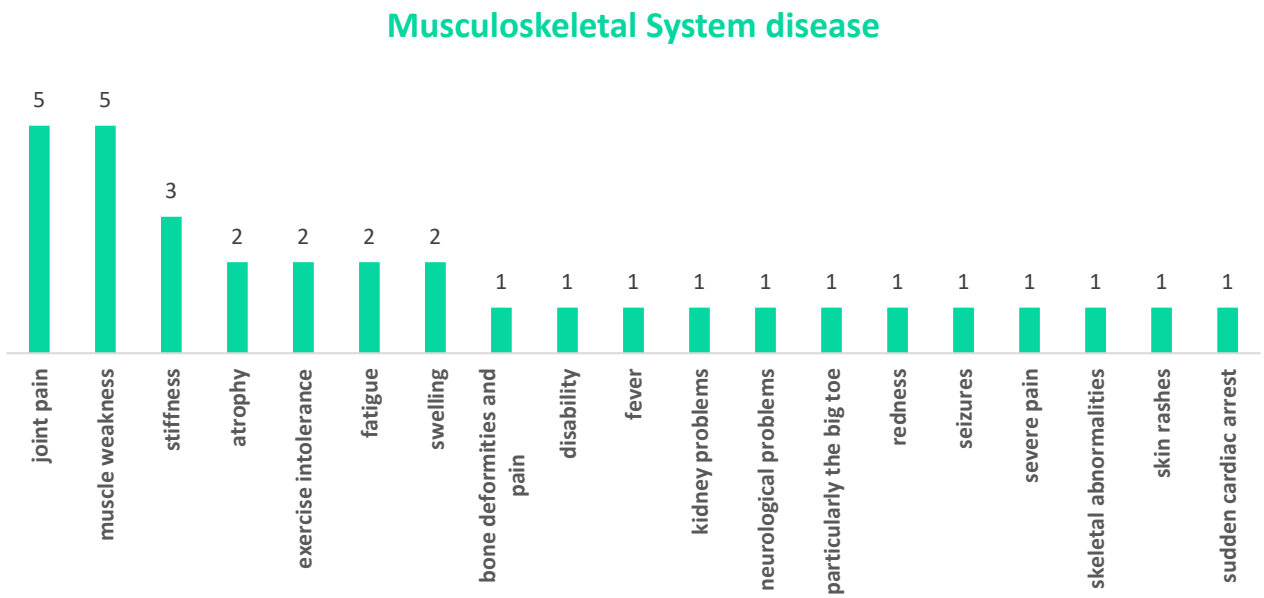
The below chart shows the symptoms frequency among the diseases in this category. Fatigue can be a common symptom of immune system diseases due to the underlying inflammation and activation of the immune system. When the immune system is activated to fight an infection or inflammation, it can release cytokines, which are chemicals that help regulate the immune response.

In autoimmune diseases, the immune system mistakenly attacks healthy tissues, leading to chronic inflammation and fatigue. Inflammation can affect various organs and tissues in the body, leading to fatigue and other symptoms specific to the affected organs.



8. Musculoskeletal System disease:

Joint pain and muscle weakness are common symptoms in many musculoskeletal system diseases, including ankylosing spondylitis, arthritis, gout, lupus erythematosus, and rheumatoid arthritis. These symptoms can be indicative of an underlying problem affecting the joints, muscles, or connective tissues.

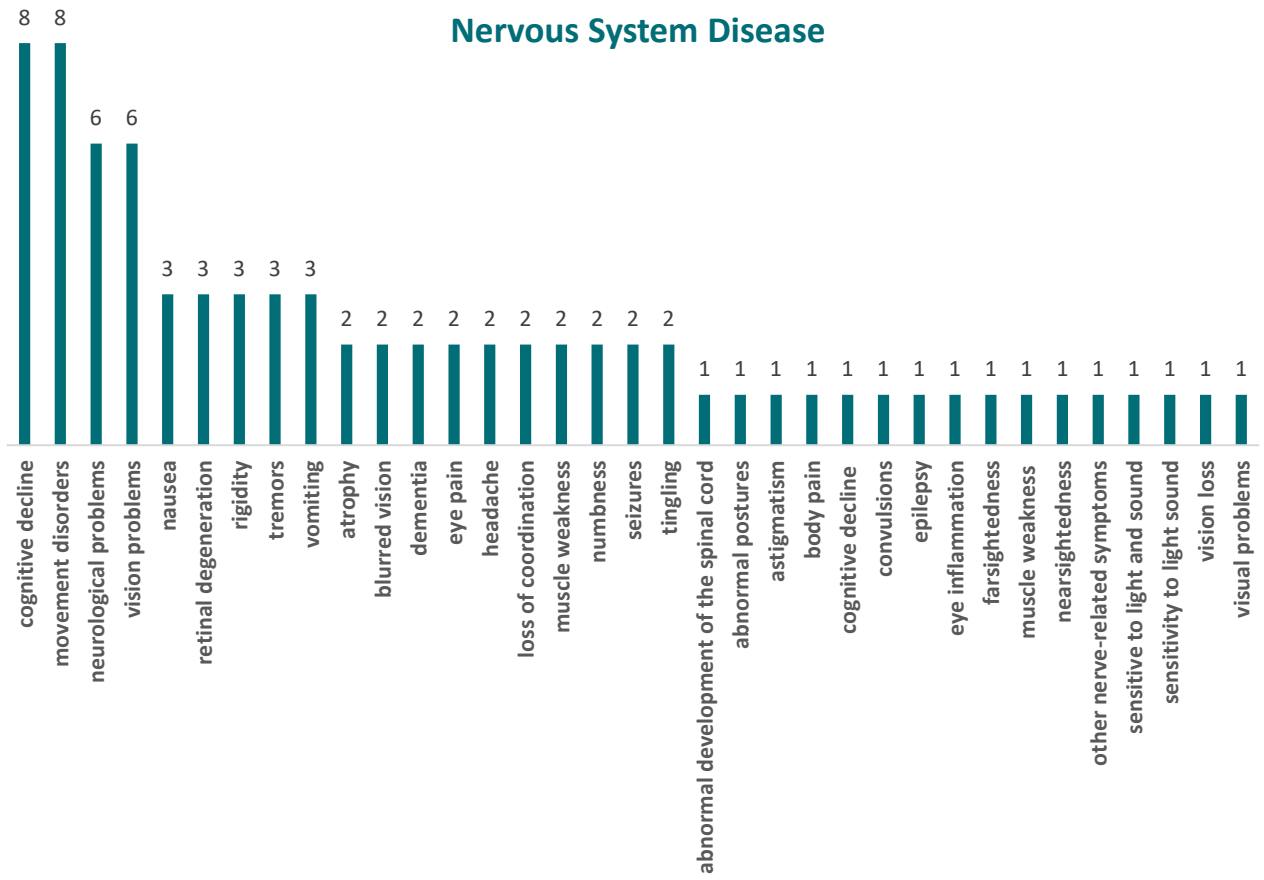


One insight you can gain from this observation is that joint pain and muscle weakness can significantly impact an individual's quality of life. Joint pain can lead to decreased mobility and difficulty performing daily tasks, while muscle weakness can affect an individual's ability to engage in physical activities and can lead to falls and injuries.

Another insight we can gain from this observation is that these symptoms can have various underlying causes. For example, in inflammatory arthritis, such as rheumatoid arthritis and lupus erythematosus, joint pain and muscle weakness can occur due to chronic inflammation of the joints and surrounding tissues. In gout, joint pain can be caused by the accumulation of uric acid crystals in the joints, while in ankylosing spondylitis, joint pain can occur due to inflammation of the spine and other joints.

9. Nervous System Disease:

Cognitive decline, movement disorders, neurological problems, and vision problems are common symptoms across various nervous system diseases, including Alzheimer's disease, Huntington's disease, multiple sclerosis, and epilepsy. These symptoms can be indicative of an underlying problem affecting the brain, spinal cord, nerves, or eyes.



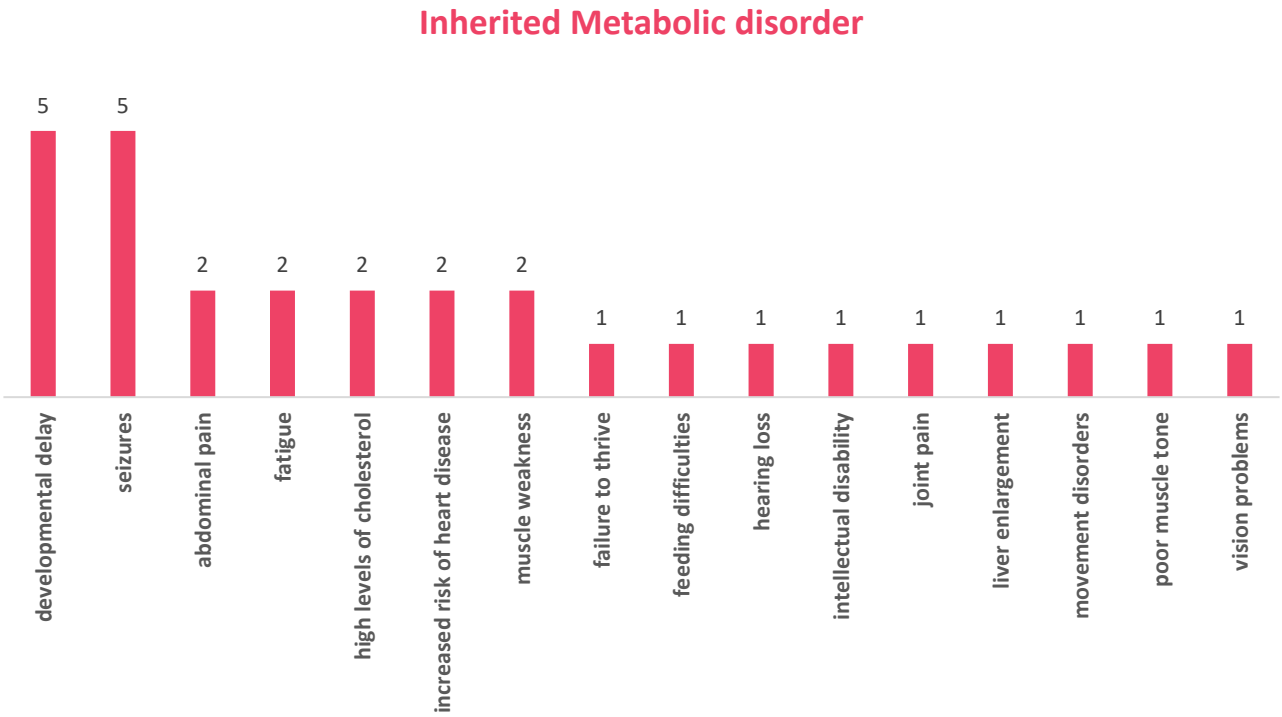
One insight we can gain from this observation is that these symptoms can significantly impact an individual's quality of life. Cognitive decline can lead to memory loss, difficulty with language, and decision-making problems. Movement disorders can affect an individual's ability to perform daily tasks, while neurological problems can cause sensory disturbances, such as numbness or tingling, or autonomic dysfunction, such as dizziness or fainting. Vision problems can affect an individual's ability to see clearly, affecting their daily activities and quality of life.

Another insight we can gain from this observation is that these symptoms can have various underlying causes. For example, in Alzheimer's disease, cognitive decline can occur due to the buildup of amyloid protein in the brain, leading to the loss of brain cells [1][2]. In multiple sclerosis, neurological problems can occur due to damage to the myelin sheath that covers nerve fibers, leading to disrupted nerve signaling [3]. In Huntington's disease, movement disorders can occur due to the degeneration of brain cells that control movement [4][5].

10. Inherited Metabolic disorder:

Developmental delay and seizures are common symptoms across various inherited metabolic disorders, including amino acid metabolism, inborn errors, cytochrome-c oxidase deficiency, hemochromatosis, hypercholesterolemia, and Leigh disease [6]. These symptoms can be indicative

of an underlying problem affecting the brain, leading to impaired cognitive function and abnormal electrical activity.



One insight we can gain from this observation is that these symptoms can significantly impact an individual's quality of life. Developmental delay can cause delayed speech, cognitive impairment, and difficulty with social interaction, while seizures can lead to injury, loss of consciousness, and impaired daily functioning.

Another insight we can gain from this observation is that these symptoms can have various underlying causes. For example, in inborn errors of metabolism, seizures can occur due to the accumulation of toxic substances in the brain, leading to damage to brain cells and abnormal electrical activity. In hemochromatosis, seizures can occur due to iron deposition in the brain, leading to brain damage and seizures. [7]

Phase 2:

Network Analysis:

We then apply various network analysis techniques to identify the key features of the network like centrality, clustering and modularity. We have created graphs based on the communities and tried to find interesting observations.

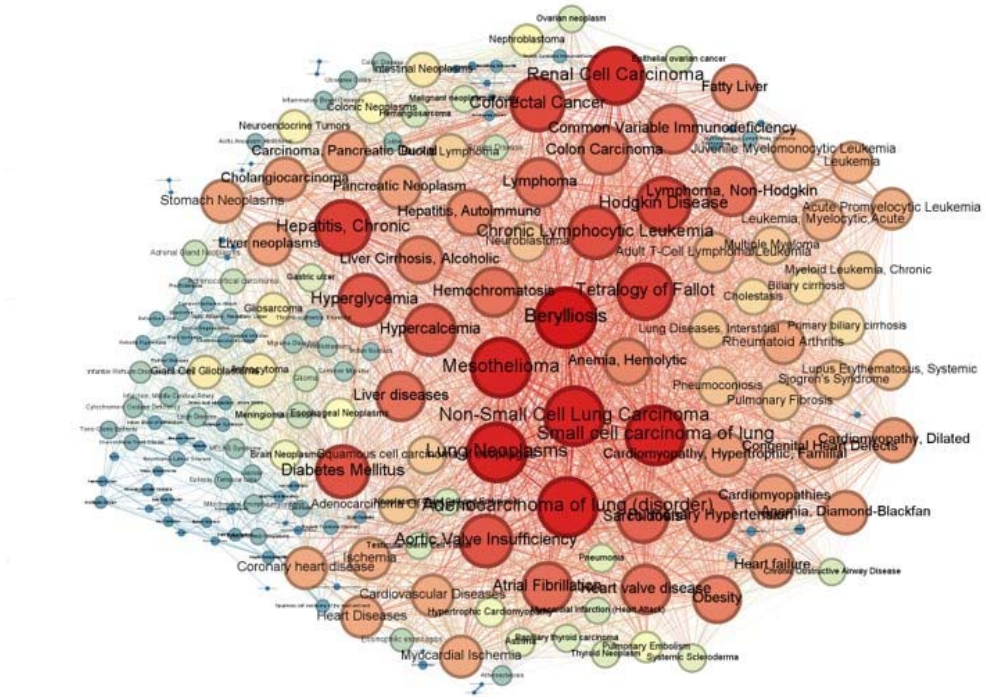
After the initial insights from the above data exploration, we applied various network analysis techniques to identify the key features of the network.

The network created based on the data collected from DoPathways dataset contains 190 nodes and 3573 edges.

1. Degree Centrality

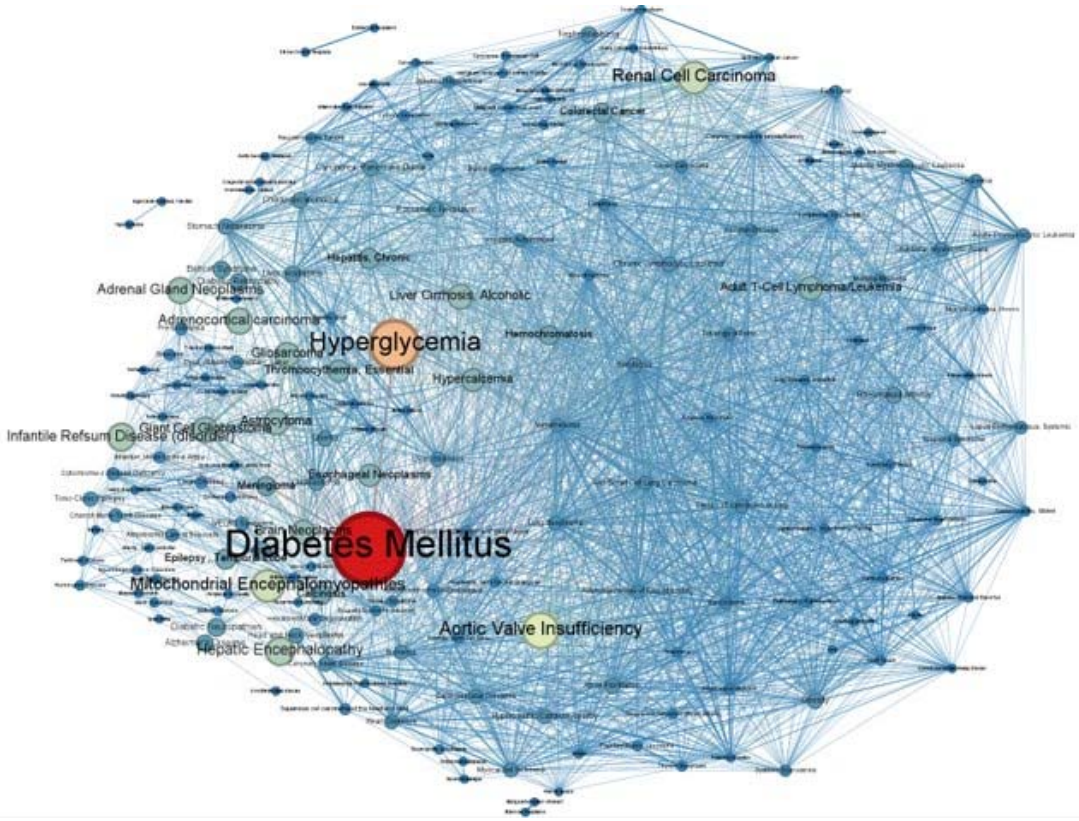
Degree centrality helps us identify the central nodes in the network. From the below graph, we can say that the nodes with larger size and reddest are the most central versus nodes with smaller size around the boundary of the network have lesser degree. The red nodes are those diseases that are highly connected to other diseases in the network. The average degree of the network is 37.611.

Using Gephi, we found that Berylliosis has the highest degree and hence, is the most central node in the network which makes it a very important node.



2. Betweenness Centrality

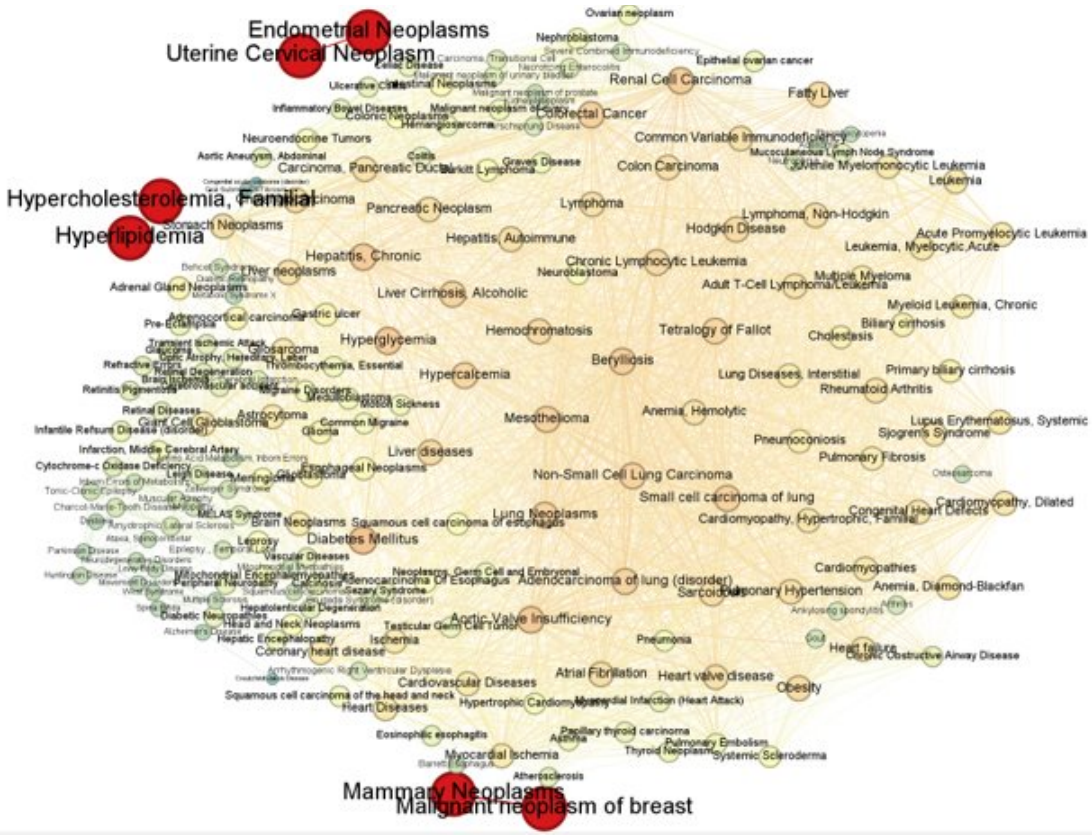
Betweenness Centrality is a metric that takes into account the number of shortest paths passing through the node. Here, we find Diabetic Mellitus has the highest betweenness centrality. Also, the network has an average shortest path length of 2.127, which also makes it one of those nodes which acts as a bridge or mediator between other nodes which helps us identify the disease sharing common symptoms or cluster of symptoms which may not be directly connected.



3. Closeness Centrality

Closeness Centrality helps in measuring a node's importance based on its proximity to other nodes. It measures how quickly information can spread through the network. In the graph below, we can see that the giant component of the network has nodes in orange having high closeness centrality. Whereas there are a few nodes in red that are disconnected from the giant component having only one edge with another node. Hence, these nodes have high closeness centrality between each other but not in consideration with the giant component.

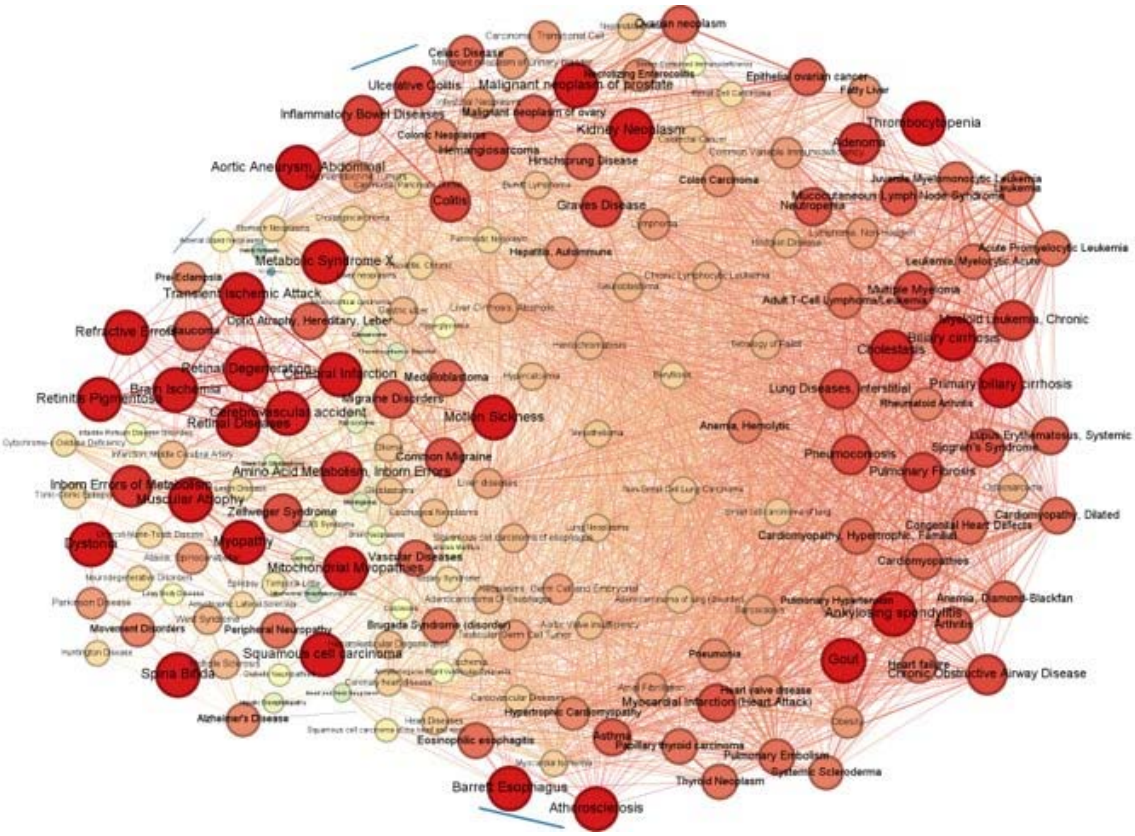
Since these orange nodes in the giant component help spread information, they act as important hubs in the network.



4. Clustering Coefficient

The clustering coefficient is a measure that helps detect the extent to which the neighbors of a node are also connected to each other. This will help us identify groups or clusters of diseases having neighbors that are highly interconnected as well.

The red nodes in the graph below depict those nodes having neighbors with a high degree centrality as well. From Gephi, we find the clustering coefficient for the network is 0.734. This suggests that the diseases in the network are highly clustered, meaning that they tend to be connected to other diseases within the same community. This can be an indication that diseases within the same community are more likely to share common symptoms and may have similar underlying mechanisms or causes.



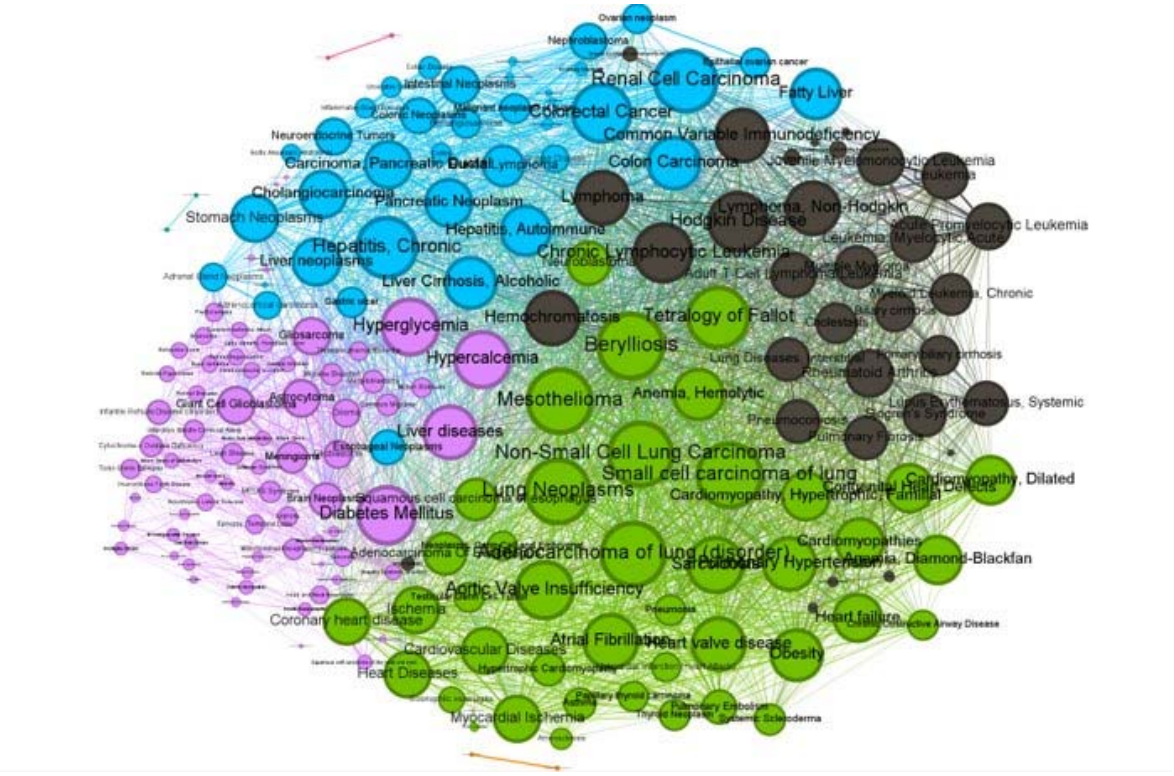
5. Community Detection

Finally, we have applied the Louvain method of modularity pre-existing in Gephi. This will help us accomplish the aim of comparing the outcomes given by the inbuilt community detection algorithm in Gephi with the actual disease categories.

The Louvain method of modularity gives a total of 7 communities in the network. Out of the 7 communities, we can see in the graph below that only 4 communities are part of the giant connected component of the network. The rest of the 3 communities are the disconnected nodes forming sparse communities.

Community detection can help us to understand if any disease could be confusing based on the community, they fall in. For example, if a disease is assigned to a community that is not consistent with its known disease category, it may indicate a potential issue in the community detection algorithm or highlight previously unknown associations between the disease and other symptoms. Therefore, community detection is a crucial tool for analyzing and interpreting the weighted symptom disease network and can help us to uncover important insights into disease associations and potential treatment strategies.

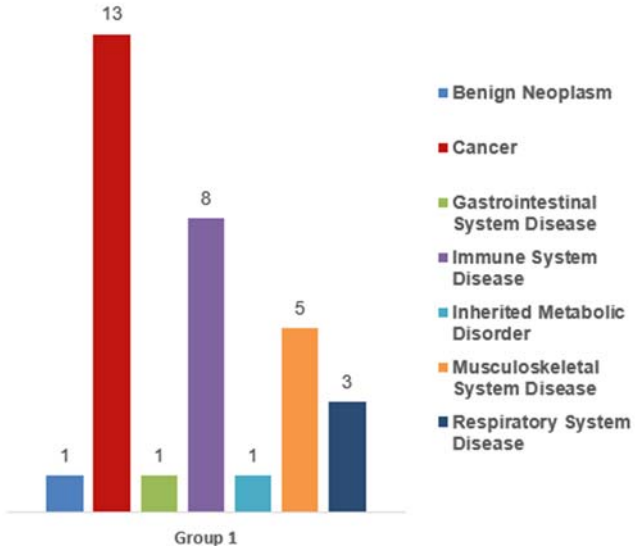
For this network, A modularity value of 0.356 at a resolution limit of 1.0 indicates that the network is moderately well-clustered, with some nodes having connections to multiple communities but still displaying a degree of separation between groups.



Results:

To understand the network better and to get interesting results from the network, we have created visualizations of the disease categories present in the major 4 groups present tin the giant component of the network.

Considering Group 1,



The above graph depicts that group 1 has the most diseases from Cancer. Based on the common symptoms shared among the diseases, the community detection algorithm in Gephi likely grouped these diseases together because they all have a potential association with immune system disorders, blood disorders, or chronic inflammatory conditions.

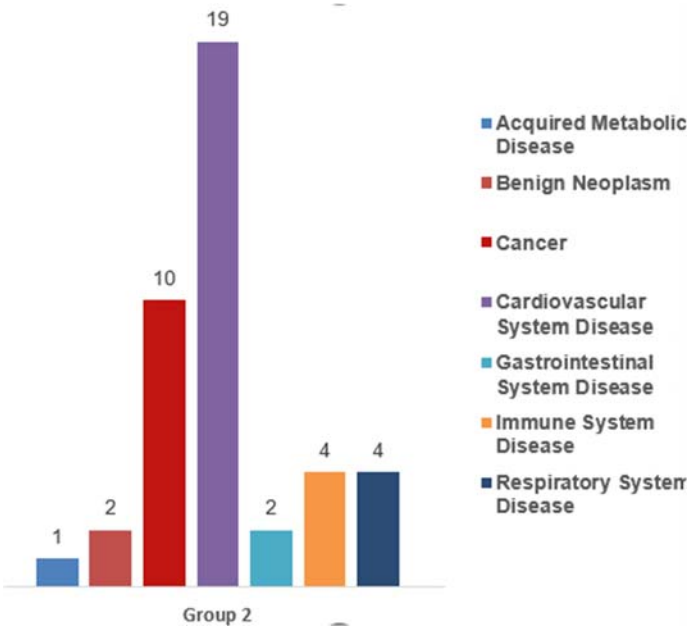
For example, several of the diseases are types of leukemia, which is a blood cancer that affects white blood cells and can be associated with immune system dysfunction. Rheumatoid arthritis and lupus erythematosus are autoimmune disorders that involve chronic inflammation and can affect multiple organs and systems in the body. Hemochromatosis and cholestasis are both conditions that can affect liver function, which is an important organ for immune system regulation and blood filtering.

Interstitial lung disease, pulmonary fibrosis, and pneumoconiosis are all respiratory diseases that can result in scarring and damage to lung tissue, leading to symptoms such as shortness of breath and coughing. These conditions can also involve chronic inflammation and immune system dysfunction in the lungs. Consider a person has Sjogren's Syndrome which as an immune system disease, this makes the person’s immune system weaker to attack the defective cells and allows cells to divide and grow developing into cancer.

People with Sjögren's syndrome have an increased risk of getting a type of cancer called lymphoma affecting the lymphatic system, a network of vessels and glands found throughout the body.[8]

Sjögren's syndrome and lymphoma are sorted into the same community even though they belong to immune system disease and cancer respectively but since they have a correlation between them which connects them to have similar symptoms resulting in them being in the same community.

Let us now consider group 2,

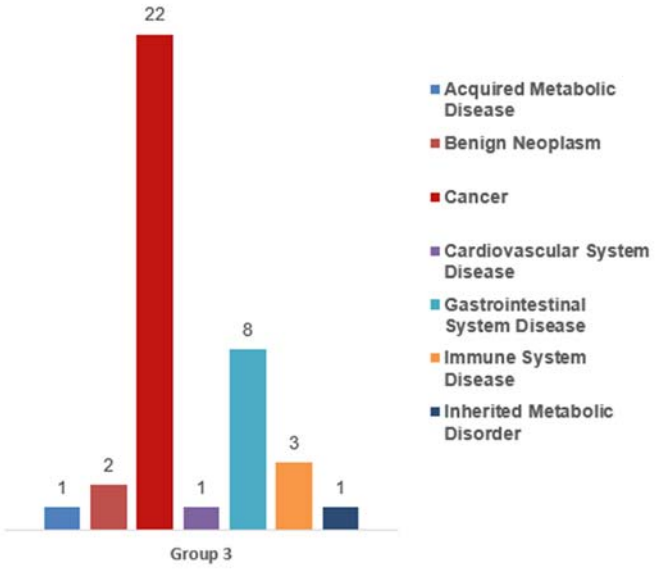


The above graph depicts group 2 having a maximum of cardiovascular diseases. We find one of the diseases from this category, i.e., Brain Ischemia has been grouped in a different community.

From the data, we can see that it has symptoms like vision problems and numbness which may have resulted in it being put with a community containing a lot of vision related diseases as well diseases having that as a common symptom.

Also, another interesting outcome is that cancer is the next highest category. So, it is seen in a recent study which says that adult survivors of cancer have a 42% greater risk of developing cardiovascular diseases like heart failure and stroke. [9]

Lets understand Group 3

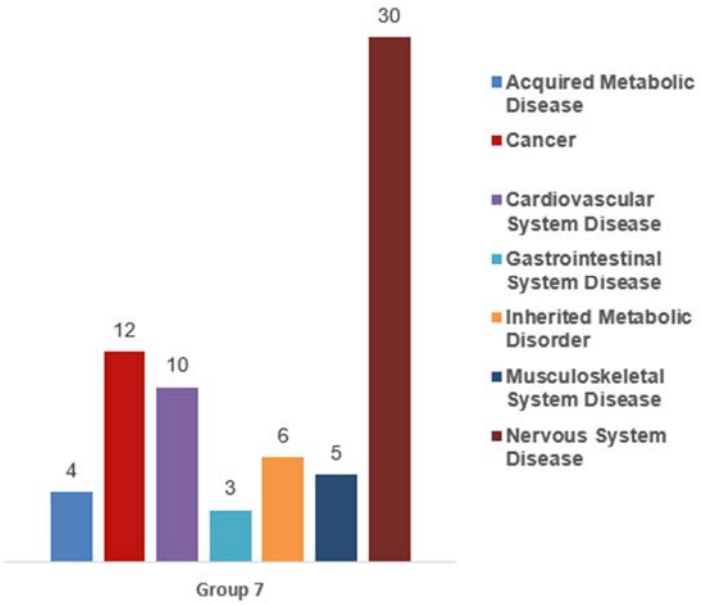


This graph talks about the third major community being cancer with next one being a gastrointestinal system disease.

We see colorectal cancer and Ulcerative Colitis being grouped together since they both share similar symptoms as well as a person suffering from ulcerative colitis having a six times greater risk of developing colorectal cancer than average risk. [10] Hence, this outcome validates this. Also chronic hepatitis can lead to liver cirrhosis and liver cancer, and inflammatory bowel disease can increase the risk of developing colorectal cancer. Celiac disease is also associated with an increased risk of certain types of cancer, such as lymphoma.

Metabolic syndrome X, which includes a group of metabolic abnormalities such as obesity, high blood pressure, and high blood sugar, has been associated with an increased risk of developing certain types of cancer, such as liver, colon, and breast cancer. Furthermore, some of the neoplasms in this community, such as pancreatic cancer and renal cell carcinoma, are known to be associated with obesity and metabolic syndrome.

Let’s understand Group 7



Lastly, this graph shows the nervous system disease to be the highest category in this community with cancer and cardiovascular disease categories following it.

Dystonia is a nervous system disease which can be associated with a side effect to head and neck neoplasms which puts them in the same community also, it can lead to strokes and heart problems as well.

This gives us an interesting outcome that all the organs in the body are highly interconnected and get affected when any part of the body may not do its work properly. Also, we see cancer as the more prominent category of disease associated with each disease. This proves that it has the greatest common symptoms with other diseases and disease categories. It can be said that on severe symptoms of diseases there is always a possibility of cancer being the issue which may not be diagnosable at that stage but can be a direction to look at if the issue isn't solved.

Discussion and Conclusion:

The present study sheds light on the intricate relationships between diseases and their associated symptoms using network analysis methods, including community detection, degree, and centrality measures. The research provides insights into disease categories that are linked to the severity of other disease categories and identifies potential areas for improvement in the community detection algorithm. These findings demonstrate the valuable role that symptom disease networks can play in identifying disease associations and predicting patient outcomes, thereby improving our ability to detect and manage diseases more effectively.

To further advance our understanding of the weighted symptoms disease network, there are several potential avenues for future research. One such avenue could be the exploration of more advanced

algorithms for community detection, which could potentially yield even more accurate and informative results. Additionally, the application of machine learning techniques could provide new insights into disease associations and patient outcomes based on the network data. Finally, the integration of other data sources, such as genetic or environmental data, could help to enrich our understanding of disease relationships and pave the way for more targeted and effective interventions.

Reference:

1. Selkoe, D. J. (2011). Alzheimer's disease. Cold Spring Harbor perspectives in biology, 3(7), a004457. https://doi.org/10.1101/cshperspect.a004457
2. Querfurth, H. W., & LaFerla, F. M. (2010). Alzheimer's disease. New England Journal of Medicine, 362(4), 329-344. https://doi.org/10.1056/NEJMra0909142
3. Frohman, E. M., Racke, M. K., & Raine, C. S. (2006). Multiple sclerosis—the plaque and its pathogenesis. New England Journal of Medicine, 354(9), 942-955. https://doi.org/10.1056/NEJMra052130
4. Ross, C. A., & Tabrizi, S. J. (2011). Huntington's disease: from molecular pathogenesis to clinical treatment. The Lancet Neurology, 10(1), 83-98. https://doi.org/10.1016/S1474-4422(10)70245-3
5. Vonsattel, J. P., & DiFiglia, M. (1998). Huntington disease. Journal of neuropathology and experimental neurology, 57(5), 369-384. https://doi.org/10.1097/00005072-199805000-00001
6. Hoytema van Konijnenburg, E.M.M., Wortmann, S.B., Koelewijn, M. *et al.* Treatable inherited metabolic disorders causing intellectual disability: 2021 review and digital app. *Orphanet J Rare Dis* **16**, 170 (2021). https://doi.org/10.1186/s13023-021-01727-2
7. Saudubray JM, Sedel F, Walter JH. Clinical approach to treatable inborn metabolic diseases: an introduction. J Inherit Metab Dis. 2006 Apr-Jun;29(2-3):261-74. doi: 10.1007/s10545-006-0358-0. PMID: 16763886.
8. https://www.nhs.uk/conditions/sjogrens-syndrome/complications/#:~:text=People%20with%20Sj%C3%B6gren's%20syndrome%20have,glands%20found%20throughout%20the%20body.
9. https://pressroom.cancer.org/releases?item=1187#:~:text=A%20recent%20study%20found%20adult,(22%25%20higher%20risk).
10. https://www.compassoncology.com/blog/how-ulcerative-colitis-affects-your-risk-for-developing-colorectal-cancer#:~:text=Ulcerative%20colitis%20patients%20have%20a,developing%20this%20type%20of%20cancer.
11. American Thoracic Society. (2017). Cough. https://www.thoracic.org/patients/patient-resources/resources/cough.pdf
12. MacNee, W. (2005). Pathogenesis of chronic obstructive pulmonary disease. Proceedings of the American Thoracic Society, 2(4), 258-266. https://doi.org/10.1513/pats.200504-035SR
13. https://github.com/dennishnf/project-symptoms-disease-network