Analyzing Disease Relationships via Hypergraph Clustering of Symptom Networks

Shyam Sankar

Abstract

The multi-way relationship between diseases and symptoms was captured through a hypergraph, where symptoms formed hyperedges that connected multiple diseases, enabling a more expressive representation of disease-symptom interactions. Unlike traditional graphs that rely on pairwise connections, hypergraphs naturally modeled higher-order relationships, allowing diseases to be clustered based on shared symptom profiles. This structure revealed implicit disease-disease connections, where two diseases were related not directly, but through multiple overlapping symptoms. By leveraging hypergraph clustering algorithms like Iteratively Reweighted Modularity Maximization and Mutual Information Maximization, the project successfully identified meaningful disease groupings, contributing to improved diagnostic decision-making and a deeper medical understanding.

1 Introduction

In the realm of medical diagnosis, understanding the intricate relationships between diseases and symptoms is a critical aspect of improving diagnostic accuracy and treatment outcomes. Traditional models typically represent diseases and symptoms as pairwise relationships, where each disease is linked to specific symptoms. However, this simplified approach fails to capture the multi-dimensional and often overlapping nature of disease-symptom interactions.[10] As medical knowledge becomes more complex, it is increasingly important to explore more advanced techniques that can reveal deeper insights into these relationships.

This project sought to address this limitation by leveraging the power of hypergraphs to model the multi-way relationships between diseases and symptoms. In a hypergraph, symptoms act as hyperedges, connecting multiple diseases simultaneously.[8] This approach allows for a more expressive and nuanced representation of disease-symptom interactions, as it captures the higher-order relationships that traditional graphs cannot. Unlike pairwise connections, hypergraphs enable a richer understanding of how diseases share symptoms, uncovering implicit disease-disease relationships that may not be immediately apparent through direct connections.

2 Related Work

[1] used Principal Component Analysis (PCA) followed by K-means clustering to identify early symptom clusters, revealing that Respiratory/Systemic cluster was linked to higher hospitalization rates, while the Nasal cluster, was linked to less severe disease and a lower risk of long-term symptoms. Symptom clusters identified using Latent Class Analysis (LCA) has been successfully used to subclassify patients with primary Sjögren's syndrome (pSS) [4]. BigCLAM, a network based clustering algorithm was used to identify 208 distinct symptom clusters across multiple diseases[5]. In conclusion, these studies have indicated the strong relationships between symptom clusters across different diseases and their specific symptom patterns, underscoring their importance in precision health and symptom management.

While experimenting with clustering algorithms, the project also focuses on developing a disease symptom hypergraph. [9] proposes a method for constructing a disease-symptom knowledge graph (DSKG) from medical web-board documents using word co-occurrence patterns, supervised learning, and PCA to identify and relate disease and symptom concepts, enabling effective cause-effect representation for non-professional diagnosis support. [6] addresses limitations of traditional bipartite graphs used in automated medical diagnosis—where diseases and symptoms are modeled as disconnected node sets—by integrating disease and symptom similarity networks to enrich the graph structure, reduce sparsity, and improve inference accuracy.
From conducting a survey of existing literature, it is evident that the use of hypergraphs in this domain remains limited. Most current approaches fail to capture the complex many-to-many interactions among symptoms and diseases that hypergraphs are well-suited to represent. Recognizing this gap, the aim of this project is to develop a hypergraph-based model for symptom-disease relationships that can more effectively encode these multifaceted connections. Additionally, the project will explore and experiment with various clustering algorithms on the hypergraph to identify patterns and groupings that could enhance diagnostic insights and support medical decision-making.

3 Disease-Symptom Hypergraph

The dataset[11] used in this study contains medical information from 4962 patient records, each comprising a diagnosed disease and a set of symptoms, encoded in a binary (one-hot) format. Each row represents an individual case, and each column indicates the presence or absence of a specific symptom. Since multiple patients can have the same disease but present with different combinations of symptoms, the same disease appears in multiple rows, capturing real-world variability in clinical presentation.
To prepare the data for hypergraph modeling, we aggregated the patient records by disease, combining all cases of the same diagnosis into a single representation. For each disease, we recorded the full set of symptoms that had ever been observed in association with it across all patients. This results in a distilled dataset where each disease is represented by a unique symptom profile. Using this structure, we constructed a hypergraph in which each node corresponds to a disease, and each hyperedge corresponds to a symptom that links all diseases known to exhibit it. This approach enables modeling higher-order relationships that go beyond simple pairwise connections, capturing the complex interdependencies between diseases that share symptoms.
This hypergraph is both novel and significant. It is constructed from real-world, clinically relevant data and introduces a fundamentally different way of modeling disease relationships—through shared symptom expressions rather than direct clinical pathways. By using symptoms to define hyperedges that connect multiple diseases, this hypergraph captures the overlapping nature of medical diagnoses in a way that traditional graphs cannot. As a result, it offers new opportunities for more accurate clustering, classification, and understanding of disease networks, making it a unique and valuable tool in computational healthcare research.

4 Methods

4.1 Iteratively Reweighted Modularity Maximization

The Iteratively Reweighted Modularity Maximization (IRMM) [3] algorithm is a method used to detect clusters (or communities) in a hypergraph. The goal of IRMM is to group together nodes that are more strongly connected to each other based on the structure of the hypergraph. It works by repeatedly adjusting how important each hyperedge is (its weight) based on how it connects different clusters and then recalculating the clustering using a well-known algorithm called Louvain modularity maximization[2].
At each step, the algorithm first builds a simpler graph (a reduced adjacency matrix) from the hypergraph, where the strength of connection between nodes reflects the current hyperedge weights. Then, it finds clusters using the Louvain method. After that, the algorithm updates the weight of each hyperedge based on how it overlaps with the discovered clusters — if a hyperedge connects many nodes from different clusters, its weight may be adjusted differently than one that stays within a single cluster. These updates continue until the weights stop changing significantly, at which point the clustering is finalized. This process helps in revealing clearer and more accurate community

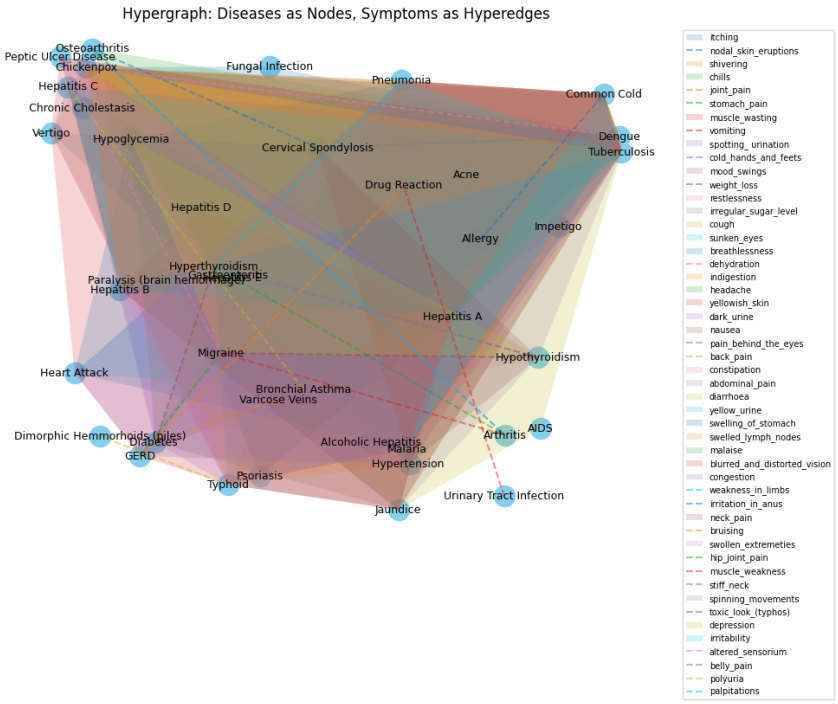

Figure 1: Disease-Symptom Hypergraph

structures in complex hypergraphs.

Algorithm 1 Iteratively Reweighted Modularity Maximization (IRMM)

Require: Hypergraph incidence matrix H, vertex degree matrix D_v, hyperedge degree matrix D_e, hyperedge weights W
Ensure: Cluster assignments `cluster_ids`, number of clusters c

Initialize weights: $W \leftarrow I$ if the hypergraph is unweighted
repeat
  Compute reduced adjacency matrix: $A \leftarrow HW(D_e - I)^{-1}H^T$
  Zero out the diagonals of A: $A \leftarrow \texttt{zero\_diag}(A)$
  `cluster_ids`, $c \leftarrow \texttt{LOUVAIN\_MOD\_MAX}(A)$
  for each hyperedge $e \in E$ **do**
    for $i \in [1, \ldots, c]$ **do**
      $C_i \leftarrow$ set of nodes in cluster i
      $k_i \leftarrow |e \cap C_i|$
    end for
    $w'(e) \leftarrow \frac{1}{m} \sum_{i=1}^{c} \frac{1}{k_i+1}(\delta(e) + c)$
    $W_{\text{prev}}(e) \leftarrow W(e)$
    $W(e) \leftarrow \frac{1}{2}(w'(e) + W_{\text{prev}}(e))$
  end for
until $\|W - W_{\text{prev}}\| <$ threshold

4.2 Mutual Information Maximization

The algorithm uses simulated annealing [7] to find an optimal clustering by minimizing entropy. It starts with random cluster assignments and iteratively proposes new assignments by changing one node's cluster. The change in entropy, Δ, determines whether the new clustering is accepted, based on the probability $\min(1, e^{-\beta(t)\Delta})$. The inverse temperature parameter $\beta(t)$ controls the

exploration-exploitation tradeoff: initially small, allowing broader exploration, and increasing over time to focus on refining the clustering and minimizing entropy.
At each step, the algorithm evaluates a proposed change by looking at how it affects entropy (computed through logZ). If the new assignment reduces entropy, it is accepted as the new current state. If it increases entropy, the algorithm may still accept it with a certain probability, especially early on—this helps avoid getting stuck in local minima. As the algorithm progresses, the acceptance probability for worse solutions drops (controlled by the cooling schedule, so it gradually shifts from exploration to fine-tuning. Throughout the process, it keeps track of the best solution seen, and finally returns that best cluster assignment.

Algorithm 2 Simulated Annealing - Mutual Information Maximization

Input: Hypergraph H, number of clusters, number of steps, annealing schedule $\beta(t)$
Output: Best cluster assignment `best_cluster`, corresponding entropy `best_entropy`
RUN_CHAINH, number_of_clusters, number_of_steps, $\beta(t)$
Initialize a random cluster assignment vector $c \in \{1, \ldots, \text{number\_of\_clusters}\}^{|V(H)|}$
`best_entropy` $\leftarrow \log Z(c)$
`best_cluster` $\leftarrow c$
$t \leftarrow 0$
while $t <$ number_of_steps **do**
  Propose $c' \leftarrow$ random neighbor of c
  $\Delta \leftarrow \log Z(c') - \log Z(c)$
  Sample $X \sim \mathcal{U}(0, 1)$
  if $X < \min\left(1, e^{-\beta(t)\cdot\Delta}\right)$ **then**
    $c \leftarrow c'$
    if $\log Z(c) <$ `best_entropy` **then**
      `best_entropy` $\leftarrow \log Z(c)$
      `best_cluster` $\leftarrow c$
    end if
  end if
  $t \leftarrow t + 1$
end while
return `best_cluster`, `best_entropy`

5 Experiments & Results

5.1 Iteratively Reweighted Modularity Maximization

Using the IRMM algorithm, the 41 diseases were partitioned into 10 clusters, showing an uneven distribution that reflects varying levels of symptom similarity. Clusters ranged from singleton groups (Clusters 3, 7, and 9) to larger, more cohesive ones like Cluster 1 and Cluster 8, each containing 7 diseases. This distribution suggests that while some diseases share many symptoms and group naturally, others have distinct profiles that set them apart. For example, Cluster 1 included mostly dermatological and infectious diseases such as Acne, Chickenpox, and Fungal Infection—conditions likely unified by symptoms like rashes or skin inflammation. Similarly, Cluster 8 grouped liver and gastrointestinal conditions (e.g., Hepatitis variants, Jaundice, Peptic Ulcer), all linked by symptoms like abdominal pain and jaundice.

The algorithm's ability to form both cohesive and isolated clusters highlights its sensitivity to symptom-based relationships. Singleton clusters such as AIDS (Cluster 3), Varicose Veins (Cluster 7), and Allergy (Cluster 9) demonstrate that some diseases have unique symptom sets not shared with others. However, clusters like Cluster 0—which combined diverse diseases such as Heart Attack, GERD, and Bronchial Asthma—raise concerns about overgrouping, possibly due to broad, non-specific symptoms like fatigue or pain dominating the structure. The network visualization in figure 2 supported these findings: tightly connected nodes appeared in Clusters 1 and 8, while Cluster 0 showed more dispersed connections. Clinical validation will be essential to further confirm the medical relevance of the identified clusters.

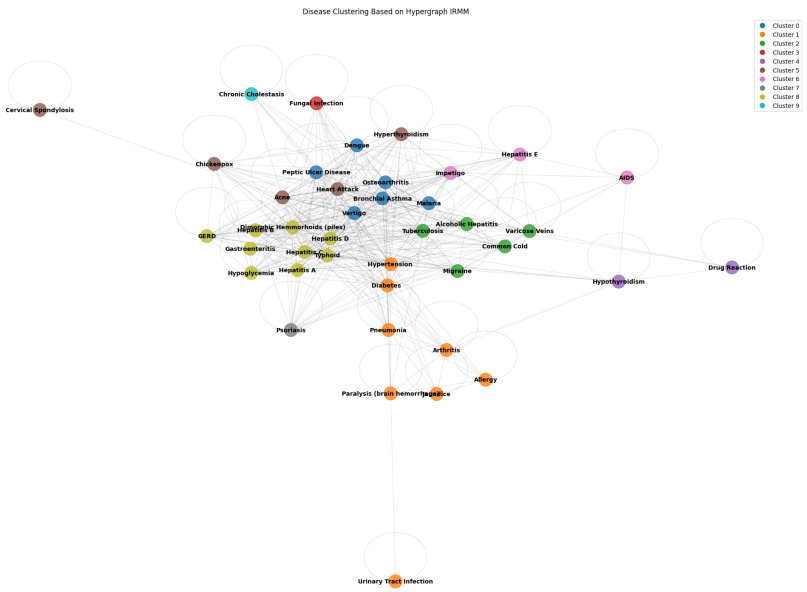

Figure 2: Result of Clustering using Mutual Information Maximization

5.2 Mutual Information Maximization

The algorithm was configured with 5 clusters and 100 iterations (after testing out different hyperparameter set and comparing best entropy values), using a linear inverse temperature schedule $\beta(t) = \beta_0 \cdot \left(\frac{t}{200} + 0.1\right)$, where $\beta_0 = 0.5$. The resulting cluster assignments are summarized in Table 1.

The algorithm assigned 41 diseases into five clusters with an uneven distribution: Cluster 0 contained 12 diseases, Cluster 1 had 3, Cluster 2 included 6, and Clusters 3 and 4 each had 10. This asymmetry reflects varying degrees of symptom overlap, with Cluster 1 emerging as the most cohesive group—containing infectious diseases like AIDS, Fungal Infection, and Gastroenteritis, which likely share symptoms such as fever and immune response disruption. Cluster 2 also primarily featured infectious conditions, though the inclusion of Hypertension points to possible overlap in general symptoms like fatigue. Clusters 3 and 4 were more diverse, spanning respiratory, digestive, hepatic, neurological, and musculoskeletal conditions. Cluster 0, the largest, grouped a wide variety of disease types, suggesting it may serve as a catch-all for diseases with broad or overlapping symptom profiles.

To evaluate clustering quality, the average Shannon entropy across hyperedges was computed, yielding a best entropy of 0.305 nats. Given that the maximum entropy for five clusters is approximately ln(5) = 1.609 nats, this result indicates a strong clustering performance—significantly better than a random baseline (typically 0.5–0.6 nats). This low entropy suggests that many symptoms were associated with tightly grouped diseases, particularly in Cluster 1. However, the broader and more heterogeneous Clusters 0, 3, and 4 may still contribute to residual entropy. A network visualization confirmed this pattern: nodes in Cluster 1 were densely connected, while other clusters appeared more dispersed. These results demonstrate the effectiveness of the method but also point to areas for improvement, such as increasing the number of optimization steps and exploring different cluster counts. Overall, the findings validate the potential of symptom-based hypergraph clustering while emphasizing the need for further refinement and domain-specific interpretation.

6 Conclusion

The Iteratively Reweighted Modularity Maximization (IRMM) and Mutual Information Maximization / Simulated Annealing (SA)algorithms were applied to cluster 41 diseases based on symptom profiles,

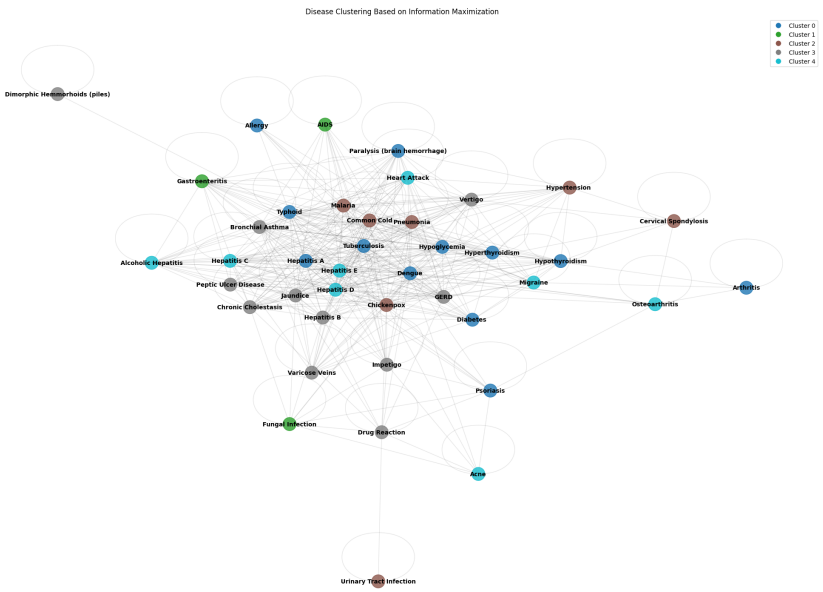

Figure 3: Result of Clustering using Iteratively Reweighted Modularity Maximization

yielding 10 and 5 clusters, respectively. IRMM produced cohesive groupings, such as dermatological/infectious diseases (Cluster 1) and liverrelated conditions (Cluster 8), alongside singleton clusters (e.g., AIDS, Allergy), demonstrating its sensitivity to distinct symptom patterns. However, heterogeneity in Cluster 0 (e.g., Heart Attack, GERD) suggests potential overgrouping, possibly due to generic symptoms or insufficient iterations. SA, with fewer clusters, grouped infectious diseases effectively (e.g., AIDS, Fungal Infection in Cluster 1) but also exhibited heterogeneity in Clusters 0 and 4, likely due to limited iterations (200 steps). Both algorithms show promise but require refinement to address overgrouping and ensure convergence to optimal solutions.

Comparing the two approaches, IRMM's modularity-based framework produced a finer granularity (10 clusters), capturing more nuanced symptom relationships, as seen in its singleton clusters, but at the cost of increased complexity. SA's entropy minimization, with 5 clusters, offered simpler groupings but struggled with diverse clusters, possibly due to a smaller solution space exploration. IRMM appears more effective for identifying unique disease profiles, while SA may benefit from increased iterations or an adjusted number of clusters to enhance coherence. Both algorithms' performance suggests that tuning parameters, such as iteration counts or cluster numbers, is critical to achieving clinically meaningful results.

Hypergraphs proved effective for symptom and disease analysis by modeling complex relationships between diseases and symptoms as vertices and hyperedges, respectively. This representation captured multi-symptom interactions, enabling both algorithms to identify clusters based on shared symptom profiles. However, the presence of generic symptoms (e.g., fatigue, pain) likely contributed to heterogeneous clusters, highlighting the need for symptom-specific weighting or feature selection to enhance specificity. With further improvement, hypergraph-based clustering holds significant potential for uncovering disease relationships, supporting diagnostic tools, and informing medical research.

References

[1] EPICC COVID-19 Cohort Study Group, Nusrat J Epsi, John H Powers, David A Lindholm, Katrin Mende, Allison Malloy, Anuradha Ganesan, Nikhil Huprikar, Tahaniyat Lalani, Alfred Smith, Rupal M Mody, Milissa U Jones, Samantha E Bazan, Rhonda E Colombo, Christopher J Colombo, Evan C Ewers, Derek T Larson, Catherine M Berjohn, Carlos J Maldonado, Paul W Blair, Josh Chenoweth, David L Saunders, Jeffrey Livezey, Ryan C Maves, Margaret Sanchez Edwards, Julia S Rozman, Mark P Simons, David R Tribble, Brian K Agan, Timothy H Burgess, and Simon D Pollett. A machine learning approach identifies distinct early-symptom cluster phenotypes which correlate with hospitalization, failure to return to activities, and prolonged covid-19 symptoms. *PLoS ONE*, 18(2):e0281272, 2023.

[2] Sayan Ghosh, Mahantesh Halappanavar, Antonino Tumeo, Ananth Kalyanaraman, Hao Lu, Daniel Chavarrià-Miranda, Arif Khan, and Assefaw Gebremedhin. Distributed louvain algorithm for graph community detection. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 885–895, 2018.

[3] Tarun Kumar, Sankaran Vaidyanathan, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, and Balaraman Ravindran. Hypergraph clustering: A modularity maximization approach, 2018.

[4] Jennifer Jooha Lee, Young Jae Park, Misun Park, Hyeon Woo Yim, Sung Hwan Park, and Seung Ki Kwok. Longitudinal analysis of symptom-based clustering in patients with primary sjogren's syndrome: a prospective cohort study with a 5-year follow-up period. *Journal of Translational Medicine*, 19(1), December 2021. Publisher Copyright: © 2021, The Author(s).

[5] Kezhi Lu, Kuo Yang, Edouard Niyongabo, Zixin Shu, Jingjing Wang, Kai Chang, Qunsheng Zou, Jiyue Jiang, Caiyan Jia, Baoyan Liu, and Xuezhong Zhou. Integrated network analysis of symptom clusters across disease conditions. *Journal of Biomedical Informatics*, 107:103482, 2020.

[6] Jingchao Ni, Hongliang Fei, Wei Fan, and Xiang Zhang. Automated medical diagnosis by ranking clusters across the symptom-disease network. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1009–1014, 2017.

[7] A.G. Nikolaev and Sheldon Jacobson. *Simulated Annealing*, volume 146, pages 1–39. 09 2010.

[8] Xavier Ouvrard. Hypergraphs: an introduction and review, 2020.

[9] Chaveevan Pechsiri and Rapepun Piriyakul. Construction of disease-symptom knowledge graph from web-board documents. *Applied Sciences*, 12(13), 2022.

[10] Fattah Muhammad Tahabi, Susan Storey, and Xiao Luo. Symptomgraph: Identifying symptom clusters from narrative clinical notes using graph clustering. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 518–527, New York, NY, USA, 2023. Association for Computing Machinery.

[11] Jay Tucker. Symbipredict, 2024.

A Appendix

Table 1: Cluster assignments for 41 diseases based on Iteratively Reweighted Modularity Maximization.

Disease	**Cluster**
AIDS	3
Acne	1
Alcoholic Hepatitis	8
Allergy	9
Arthritis	4
Bronchial Asthma	0
Cervical Spondylosis	6
Chickenpox	1
Chronic Cholestasis	8
Common Cold	0
Dengue	1
Diabetes	2
Dimorphic Hemmorhoids (piles)	5
Drug Reaction	1
Fungal Infection	1
GERD	0
Gastroenteritis	5
Heart Attack	0
Hepatitis A	8
Hepatitis B	8
Hepatitis C	8
Hepatitis D	8
Hepatitis E	8
Hypertension	6
Hyperthyroidism	2
Hypoglycemia	2
Hypothyroidism	2
Impetigo	1
Jaundice	8
Malaria	5
Migraine	2
Osteoarthritis	4
Paralysis (brain hemorrhage)	5
Peptic Ulcer Disease	8
Pneumonia	0
Psoriasis	1
Tuberculosis	0
Typhoid	5
Urinary Tract Infection	1
Varicose Veins	7
Vertigo	6

Table 2: Cluster assignments for 41 diseases based on simulated annealing algorithm for mtutual information maximization.

Disease	**Cluster**
AIDS	1
Acne	4
Alcoholic Hepatitis	4
Allergy	0
Arthritis	0
Bronchial Asthma	3
Cervical Spondylosis	2
Chickenpox	2
Chronic Cholestasis	3
Common Cold	2
Dengue	0
Diabetes	0
Dimorphic Hemmorhoids (piles)	3
Drug Reaction	3
Fungal Infection	1
GERD	3
Gastroenteritis	1
Heart Attack	4
Hepatitis A	0
Hepatitis B	3
Hepatitis C	4
Hepatitis D	4
Hepatitis E	4
Hypertension	2
Hyperthyroidism	0
Hypoglycemia	0
Hypothyroidism	0
Impetigo	3
Jaundice	3
Malaria	2
Migraine	4
Osteoarthritis	4
Paralysis (brain hemorrhage)	0
Peptic Ulcer Disease	3
Pneumonia	2
Psoriasis	0
Tuberculosis	0
Typhoid	0
Urinary Tract Infection	2
Varicose Veins	3
Vertigo	3