Understanding Long Covid-19 Patterns in Pediatric Patients using Network Analytics

by

Ornela Hogu

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 7, 2022

Keywords: Bipartite graph, projections, COVID-19, long COVID, Disease Network, Pediatric Patients

Approved by

Xiao Qin, Alumni Professor, Computer Science and Software Engineering
Ashish Gupta, Co-chair, Professor of Business Analytics, Department of Systems and Technology
Pankush Kalgotra, Assistant Professor, Department of Systems and Technology

Abstract

COVID-19 has had a long-term impact on the quality of life, work, and society by manifesting in the form of Long COVID. Long COVID is relatively less understood condition due to its recency. The challenge is even greater among pediatric population, which represents 19% of overall Long COVID cases, due to lack of research and ample data. We study three main research questions in the study. First, what are the most frequently occurring chronic conditions among pediatric patients suffering from Long COVID at different time periods? Second, what are the most frequent non-chronic conditions among pediatric patients suffering from long-Covid across different age segments? Third, what are various clusters of chronic and non-chronic conditions that exist among pediatric patients diagnosed with Long-COVID.

Using N3C (National COVID Cohort Collaboration) data, we analyze health records of ~500K pediatric patients suffering from long COVID across 72 different sites. We apply network analytics approaches to model various chronic and non-chronic conditions that pre-exists in patients diagnosed with Long COVID. In the first part, we model two network types to capture the chronic and non-chronic diseases in pre and long Covid. In the second part, created bipartite graphs and its projections to generate network clusters of pre-existing diseases and coexisting disease network for Long COVID. We then applied two community detection algorithms, Louvain and Leiden algorithms, on these projections to identify clustering patterns of diseases. We analyzed and interpreted top clusters and observed a high dominance on the conditions related to the pregnancy, neoplasm(cancer), infectious, parasitic diseases and other categories. To develop insights into co-existing non-chronic conditions, we segmented the data across three pediatric age groups (0-4, 5-11, 12-17 years).

Our findings suggest that Long COVID co-exists with four highly frequent chronic conditions, namely, asthma, anxiety, obesity, and lipoprotein metabolism disorders. For all pediatric patients suffering from Long COVID, we found five dominant non-chronic co-existing

conditions: acute upper respiratory infections, fever, Acute pharyngitis, deficit hyperactivity disorders, and cough. However, we observed some unique conditions when segmented across different age groups. For example, sleep disorders and severe stress were dominant across 11-17 age group. Using Louvain Community detection algorithm, we identified five key clusters. For example, cluster one (approx. 14% of data) had higher levels of teen pregnancies, infectious or parasitic diseases, and relatively lower levels of mental or behavioral disorders while cluster two (approx. 5.5%) had higher instances of neoplasms, and infectious or parasitic diseases. Our findings have important implications for pediatric care providers and researchers. Using network analytic approaches, we identified various clusters of chronic and non-chronic conditions that exist with Long COVID diagnosis among pediatric population. Such an understanding could provide early insights into the nature of pediatric patients who are likely to develop Long COVID from COVID-19.

Acknowledgments

I would like to express my sincere gratitude to my research supervisor Dr. Gupta. His guidance, timely advice, and scientific approach carried me through all the phases of this study. I would like to thank my committee members Dr.Qin and Dr.Kalgotra for their recommendations, comments and support throughout this journey.

I would like to give special thanks to my parents and my sister for the continuous support and understanding while completing my master's degree and undertaking this research. Being far away from home during the unprecedented times of Covid-19 was very challenging.

A special thanks goes to the Fulbright Program – a prominent and competitive scholarship that fully funded my studies in the U.S. This program has provided to many opportunities to build connections both personal and professional and be an ambassador of my own country, Albania.

List of Abbreviations

CDC – Center for Disease Control and Prevention

EMR – Electronic Medical Records

NC – Normal COVID

PC- pre- COVID

LC – long COVID

ICD –10 CM- The International Classification of Diseases, Tenth Revision, Clinical Modification

ICD –10The International Classification of Diseases, Tenth Revision

PCPP- Positive Covid Pediatric Patients

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

Although the outbreak of COVID-19 found the world unprepared to face a pandemic, it has shown how much science has advanced and developed in terms of biomedical/pharmaceutics and big data. The implementation of EMR (Electronic Medical Records) in the healthcare setting has a crucial importance in building networks and predictive models, because it has established an environment to store historical data, treatments, and other information that are essential. The historical data of patients along with COVID-19 tests (including PCT and images), have paved the path to the use statistical and the state-of-art of Artificial Intelligence/Machine Learning techniques to project the number of cases, early diagnose high-risk diseases, predict long term effects and so on. The necessity to leverage the impact of the pandemic has become an intriguing topic for data scientists, researchers, and machine learning engineers. Data engineering, artificial intelligence, and data visualization are being widely used to keep track, analyze, diagnose, and predict the trend of COVID-19 cases.

In the early COVID phases, children, and young people (0-18years old) appeared to be resilient and invulnerable compared to other groups. However, recent studies and data collected have shown that this domain is fragile not only from COVID-19, but also suffering from long-COVID symptoms which include “brain fog”- cognitive impairment, headaches, dizziness, cough, chest pain, fever, shortness of breath, among others. A complete list of post-COVID symptoms is available in the CDC [1] official site.

As of January 21, the American Academy of Pediatrics [2] has reported 1 million cases, which is 4 times the rate of the peak of last winter. In addition, other findings show that 17.8% of

[1] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/post-covid-conditions.html
[2] https://www.aap.org/en/pages/2019-novel-coronavirus-covid-19-infections/children-and-covid-19-state-level-data-report/

positive cases represent children, and the number of hospitalizations has increased. All these observations indicate the vulnerability and sensitivity of the pediatric domain. The dynamic and variations in the pediatric patients might be related to the exposure and other host factors. The shutdown of schools and kindergartens kept them isolated, and there were fewer cases reported. Then, bringing them back to pre-schools and kindergartens resulted in an increase in number. Researchers are actively studying and evaluating demographics, clinical and laboratory results by feeding various machine learning and deep learning models to prevent and predict various upcoming events. The focus of our research is related to the usage of big data in building recommendation models to identify patterns and correlation among diseases from pediatric patients suffering long-term effects of COVID. According to CDC official site, **post-COVID conditions** are a wide range of new, returning, or ongoing health problems people can experience **four or more weeks** after first being infected with the virus that causes COVID-19. The terms long COVID, long-haul COVID, post-acute COVID-19, long-term effects of COVID, or chronic COVID are used interchangeably to address the same concept. We focus to answer the following questions:

1 What are the most frequently occurring chronic conditions among pediatric patients suffering from long COVID at different time periods?
2 What are the most frequent non-chronic conditions among pediatric patients suffering from long-Covid across different age segments?
3 What are various clusters of chronic and non-chronic conditions that exist among pediatric patients diagnosed with long COVID?

In addition, our purpose is to overcome some of the challenges and limitations related to the findings in pediatric patients because of lack of big and real data. An explanation is provided in the literature review chapter.

1.2 Thesis Organization

This thesis conducts the study of long COVID in the pediatric patients. Chapter 2 presents the literature review focusing in three pillars: analyzing studies in the pediatric domain, long COVID symptoms in this domain and disease networks modeled as bipartite graphs. The work presented in chapter 3 explains and interprets insight from our dataset. It provides a high-level workflow of methodology used. We explain two concepts: time modeling for long COVID and the importance of the ICD-10 CM codes. In Chapter 4, we have defined and designed two networks: Chronic – Diseases Network and Complete- Diseases Network, which is based on chronic-nonchronic diseases in pre-COVID and nonchronic diseases in long COVID. These results are summarized in a tabular fashion and visualized with Sankey diagrams. Then, we defined Complete- Diseases Network as a bipartite graph and obtained the projections in pre and long COVID. We evaluated these projections by using two community detection algorithms: Louvain and Leiden algorithms. The top-communities obtained were analyzed and compared. Finally, Chapter 5 provides a summary of the work accomplished, limitations and future work directions.

2 Literature Review

2.1 Analyzing Studies in Pediatric Patients' Domain

In the review [1], they analyzed 129 studies from 31 countries comprising 10,251 children of which 57.4% were hospitalized. They recommended that children predominantly faced mild form of infection, but they are at risk of more severe outcomes. Their analysis presents a comparison of clinical symptoms, management, and outcomes among reported pediatric patients. The criteria for the severity of a disease were defined within each individual study considering parameters like admission to intensive care (ICU), usage of ventilation, multiorgan failure and the presence of hypoxia. Establishing the severity parameter with a quantitative value is an exceedingly challenging task because it is more of a qualitative attribute. Numerous studies define it differently, making the evaluation and comparison among them difficult to represent a meaningful and consistent outcome.

In another statistical analysis [2] evaluated the PCR tests from 3118 pediatric patients to assess the epidemiological, demographic, clinical characteristics, and laboratory findings of pediatric patients. The data was collected from a hospital in Ankara, Turkey. They reported a positive test rate of 19.9 % and a mortality rate of 0.32%. In addition, they compared with the outcome of positive test rates in Texas (7.3%) and during the first peak in England (4%). In overall, they concluded that COVID-19 caused a wide spectrum of symptoms in pediatric patients, mostly mild clinical presentation, but they emphasized that more severe conditions may be in children with early age and comorbidity.

A lot of research is conducted and summarized in the comprehensive review by [3]. The authors observed that most of the models providing good prediction accuracies used image datasets, such as X-rays or CT-scans to report the results. However, the approach for analyzing chest CT-images in pediatric patients suffers from two main disadvantages: cost and risk of

developing cancer due to radiation exposure. In addition, an important concern is related to confidentiality and privacy of personal medical records, especially accessing data for patients under 18 years old. Last but not least, it is crucial to use real data to avoid biased results when applying predictive models. Most of the studies have been using public repositories or synthetic data which does not guarantee the uniqueness of the records and the quality of them. Moreover, sometimes the number of records gathered is not enough to train and test predictive models, leading to poor and inaccurate outcomes.

2.2 Long COVID Symptoms in Pediatric Patients

There is a significant paucity of research on understanding the long-Covid symptoms in pediatric patients and estimating children that are more likely to have a propensity for developing long Covid symptoms. [4] machine learning techniques on pediatric cases with COVID-19 infections to predict the results of CR scans by using clinical laboratory data and RT-PCR positive results. From the computational perspective, the number of scans represented a small sample size of 200, which could result in biases conclusions when training and testing machine learning models. As mentioned above, the usage of image data for pediatric patients has some leak points. [5] reported a case study of five Swedish children aged 9–15 who had experienced symptoms for more than 2 months after clinical diagnoses of COVID-19, in which females appeared to be overrepresented. All had fatigue, dyspnea, heart palpitations or chest pain, and four had headaches, difficulties concentrating, muscle weakness, dizziness and sore throats. Another study,[6] based on 58 children and adolescents reported to suffer long Covid symptoms. These symptoms included fatigue in 12 (21%), shortness of breath in 7 (12%), exercise intolerance in 7 (12%), weakness in 6 (10%), and walking intolerance in 5 (9%) individuals. Older age, muscle pain on admission, and intensive care unit admission were significantly associated with long COVID.

A very intriguing outcome by [7], they observed an increased rate of Type 1 diabetes among US children during COVID-19. The lack of social activities appeared to cause long term adiposity at children in Singapore, one year after the lockdown [8].

2.3 Disease Networks using Bipartite Graphs

Bipartite networks are important to design complex systems in the real world. Many real-domains such as patient-disease, disease-genes, author-paper, customer-product are bipartite networks in nature. Therefore, our aim is to identify and detect communities that provide meaningful information and outcomes. The concept of communities is very frequent in social networks, biological networks, and so on. Communities play a crucial role in understating human diseases, and they cannot be explained only on degree distribution, but it is connected to the fact of who connects to whom [9] . We need algorithms to identify communities because the Bell number is not an efficient strategy when dealing with millions of nodes in a network. In the book [9], explains two algorithms used to detect communities: The Ravasz algorithm - an agglomerative algorithm, which uses the average cluster similarity and the Girvan-Newman algorithm, a divisive algorithm. These two techniques led to a hierarchical tree, called dendrogram which is further optimized by using modularity in order to decide an optimal cut for large networks. As a result of this optimization, the Louvain algorithm is applied to identify communities in large networks, along with Leiden and other versions. However, in the human diseases network, overlapping communities are very frequent, because practically a node belongs to more than a single community. Therefore, to identify these overlapping communities there exists two algorithms: the clique percolation method and link clustering algorithm.

Our study consists of massive data and a dynamic network over time. In this setting, patient-disease and disease-disease associations generated in various timeframes are extremely large. It is impossible to read and interpret any result when more than nodes the number of edges is

very dense. Some research used the concept of tripartite graphs to monitor and detect useful information related to Covid. However, the implementation was based on the social network, twitter data [10]that is quite different from our domain of data. Instead, [11], proposed an improved bipartite network projection method to detect metabolite-disease associations based on linear neighborhood similarity. KATZ model and Bipartite Network Recommendation Algorithm (KATZBNRA) were applied in much research to discover potential association like micro-disease [12], metabolite-disease [11], CircRNA- disease [13], [14]. In our study, the methodology is unique and adapted to the environment where the data is stored.

3 Dataset and Preprocessing

Figure 3.1 provides a high-level workflow of the steps used to conduct this research. The sections: 3.1, 3.2 and 3.3 provide details related to the first two blocks. In chapter 4 are explained and developed the remaining steps (blocks).

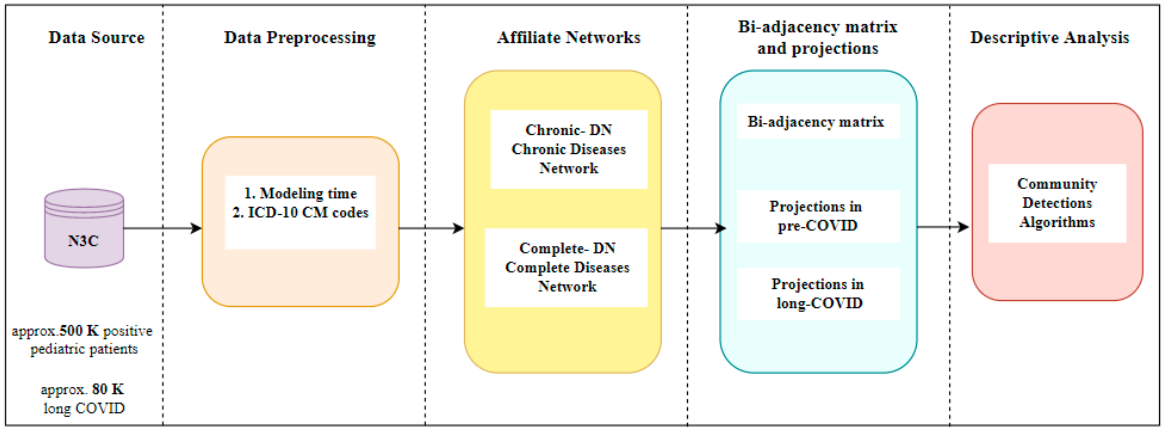


Figure 3.1: Methodology: *High-Level Workflow*

3.1 Pediatric dataset

This research is based on the data obtained from "The National COVID Cohort Collaborative (N3C)."[3] It is a secure platform that stores clinical and demographics data for patients tested for or diagnosed with COVID-19. It provides real data information from multiple sites which allows us to generate reliable and pragmatic outcomes that represent a diverse and a wide range of populations. It consists of 72 sites and approximately 4.9 M positive patients. The definition of positive patients is based on: Positive Lab Measurements (PCR, Antigen, or Antibody) or Covid diagnosis. In table 3.1. we have extracted some statistical data related to the pediatric population. In appendix A.1, an ER diagram explains the detailed steps we used to extract this information. We identified the number of positive pediatric patients reported is ~500K (16%

[3] The analyses described in this thesis were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data and the organizations and scientists who have contributed to the on-going development of this community resource.

of cases among all positive cases). This percentage is very close to the one reported by the American Academy of Pediatrics. In addition, the number of positive cases in terms of pre-COVID is ~434K and in long COVID ~80K. Lastly, we can observe that the distribution of positive cases among females and males is the approximately the same in PC and LC. In addition, the number of records where the gender is not recorded (No matching concept) is exceedingly small (~0.09%) in LC, which does not impact the overall distribution. Hence, in the pediatric population the age parameter is interesting and important to use for further estimation when studying distinct groups.

Pediatric Patients		**pre-COVID**	**long COVID**
		total = *434,302*	**total =***83,031*
Age	Mean ± SD	9.6 ± 5	9.3 ± 5.2
	Min	0	0
	Max	17	17
	Median	10	9
Gender	Male	222k (~51.23%)	42,4k (~51.9%)
	Female	212k (~48.77%)	39,9k (~48%)
	No matching concepts	-	749(~0.09%)

Table 3.1: Demographic and statistical outcomes among positive pediatric COVID cases

3.2 Modeling Time for Long COVID

In the literature review, many studies have used a period that includes a range of a couple of months up to one year to evaluate COVID-19 symptoms and the effects in terms of long COVID. Furthermore, many studies had no clear definition or reference of the time window used in their evaluation. In our study, as previously stated, we are referring to the CDC definition of long COVID. The patients with symptoms/diseases/conditions that persist from

four weeks after being diagnosed with COVID-19 will be considered long COVID cases. This leads to the interpretation that the period from the day of being diagnosed up to four weeks will be considered as a normal condition or normal COVID (NC). The steps performed to preprocess our massive and dynamic data are based on the timeframe Figure 3.2. Our approach was to split the time in three main events: pre-COVID, normal COVID, and long COVID.

Initially, we started off by using standard tables to extract all positive pediatric patients, their diseases, and other details related to the start and end date of being diagnosed with a specific disease. Among positive pediatric patients, we further proceed on dividing in three main timeframes:

1. Pre-COVID (PC1)- we looked back up to one year from the date of being identified as a positive case. In this table, we stored all pediatric patients identified as a positive case- and diseases they had before the date of T_{Covid}
2. Normal COVID (NC) – we extracted and stored all the pediatric patients, and their diseases from the day they were identified as positive cases for up to four weeks.
3. Long COVID (LC)- as mentioned before, the term Long COVID relates to the period of four or more weeks after being identified as a positive case. Cutting off up to one year is related to the fact that we are dealing with big data, and we had to overcome the limitations of time preprocessing.

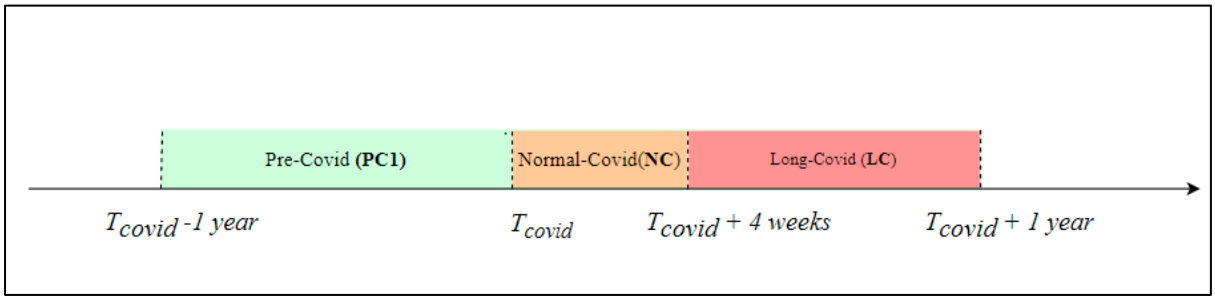


Figure 3.2: Modeling Time for Long COVID

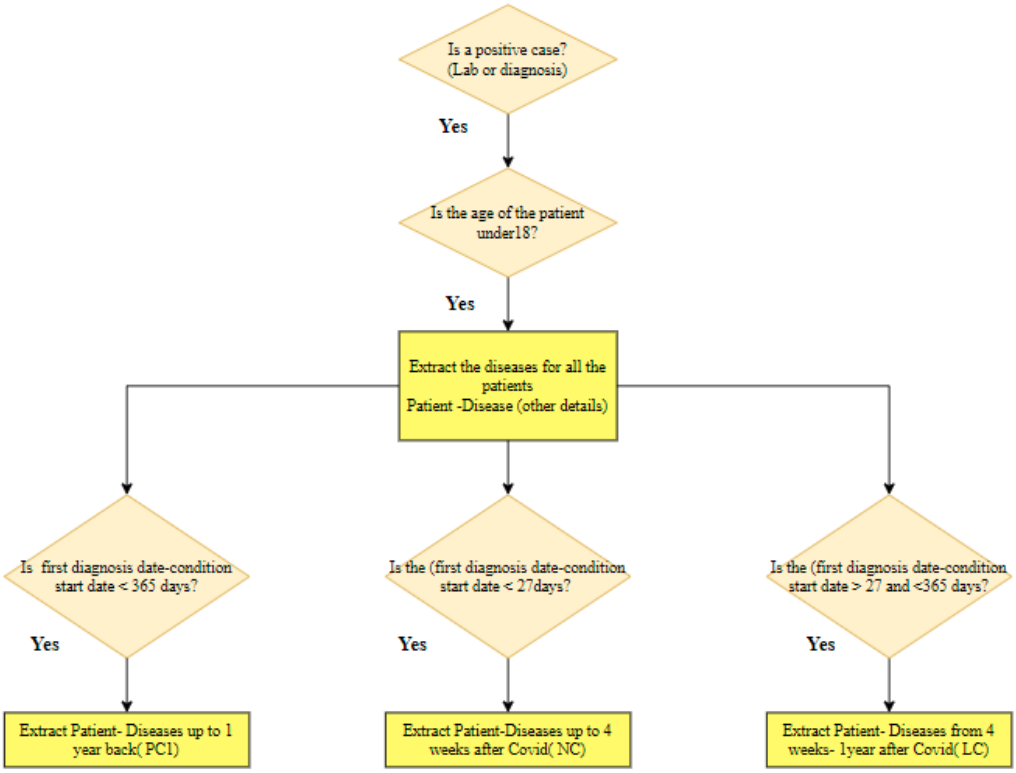


Figure 3.3: High-level workflow (Positive Patients-Diseases) in the time window

The workflow diagram in the figure 3.3, presents the logic applied to extract and generate three master lists in: pre COVID, normal COVID, and long COVID. In appendix A.1, we have provided an ER diagram of the standard and custom tables used.

3.3 ICD-10 CM Codes Importance

ICD 10 code (International Classification of Diseases, Tenth Revision, Clinical Modification) is a classification system used globally to represent diseases and conditions, health problems, symptoms, and so on. One main usage is to manage claim reporting and payment services. For more than a century, the International Classification of Diseases (ICD) has been the basis for comparable statistics on causes of mortality and morbidity between places and over time.[4]

[4] https://www.who.int/standards/classifications/classification-of-diseases

In healthcare, another standardized system is SNOMED which stands for Systematized Nomenclature of Medicine -- Clinical Terms. It provides too many details for the terms and inputs in the EHR. The ICD-10 CM codes are more efficient to capture diseases [15]. By using this system, we would have had 49 thousand diseases to consider, which would not have helped to our purpose of reducing the state representation of diseases. Therefore, we evaluated that by using the ICD-10 CM codes in terms of category level we would reduce our disease space at 1945 diseases. The heuristic we applied is based on extracting 3 digits' part from the ICD-10 CM codes. The ICD-10 CM code format is provided in Figure 3.4. below:

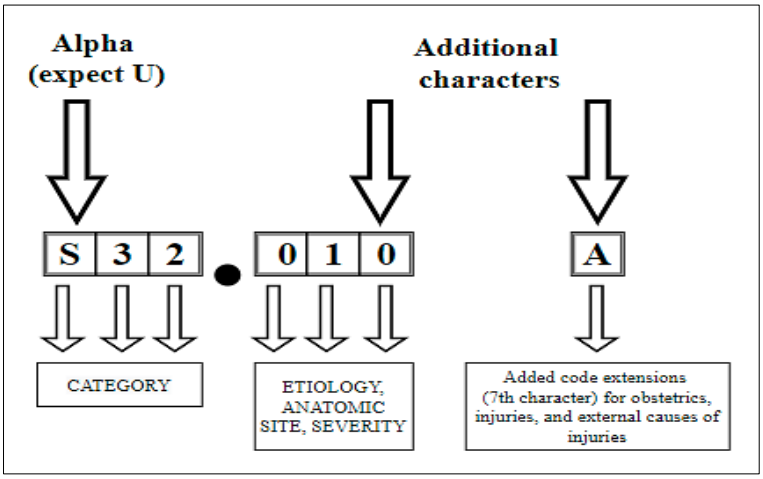


Figure 3.4: ICD-10 CM code Format

ICD-10 CM code consists of 3 or 7 characters, where:

1. The first three characters of an ICD-10 code describe the general type of a disease, condition, or injury. It is called a category, and it starts with a letter (all letters can be used besides U) followed by two numbers. The part after the decimal point is the subcategory. The characters from 3rd to 7th can be alpha or numeric.
2. Subcategory- is the part after the decimal point. It provides two subcategories to further inform on details of the disease.

3. The last character is an extension used to record if it is the first time a health care provider has seen the patient for this disease/injury/ condition, it is labeled as the initial encounter.

In figure 3.5. is shown a simple example of the ICD 10 code [5].

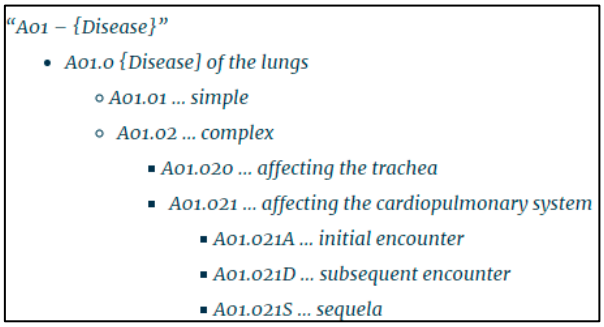


Figure 3.5: ICD-10 CM code theoretical structure

In figure 3.6. we see a practical example of the ICD-10 code of the J45-Asthma. Here, we can observe that the code J45 Describes the category and J45.2 gives a detailed explanation of the severity of the disease. From this practical example, we realize that by extracting the first three characters from the ICD-10 CM code, we are not losing meaningful information. This means that if patients that have 5 diseases that fall under the category of J45, we will consider only one. Undoubtedly, this heuristic is a practical and efficient approach to reduce our data dimension. Refer to appendix A.2 for the detailed steps to generate 3 digits from ICD-10 code for all the patients and their diseases.

[5] https://www.medicalbillingandcoding.org/icd-10-cm/#:~:text=ICD%2D10%2DCM%20is%20a,decimal%20point%20and%20the%20subcategory.

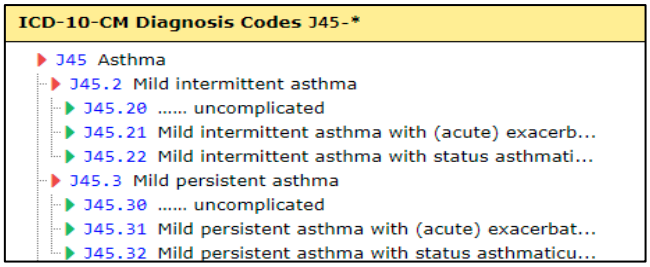


Figure 3.6: ICD-10 code example for J45- Asthma[6]

In our dataset, the ICD-10 code is not recorded and stored in an exact waypoint. For instance, for a particular code we might have other information that is not related to the ICD 10 code like *ICD-10CM:R41.82.* In this case, we had to parse and capture only the R41. We had to manage various cases that contain special characters or texts to extract only 3 digits, the category fraction. In the final results, we had some codes such as: F76, F37, E82 etc, which are not found in the ICD-10 codes. The reason is because in the system there are codes like E821.9 (an ICD-9 format) and this pattern is impossible to manage when we had to extract 3 digits' parts. However, these are only few cases that comes as a result of dynamic and challenging environment of the healthcare. In appendix A.3, we have provided an ER to explain the steps to generate a master list of patients-diseases-ICD codes.

At this point, why did we use ICD-10 format and not ICD-9 or ICD-11? An evaluation was made and related to the fact that ICD-9 code was effective up to October 1, 2015. [7] On the other hand, the latest version ICD-11 started to be effective by January 2022.[8] These versions are being used not in the time window that we are evaluating. All patients experiencing COVID-19 started to appear in the HER systems in March 2020 and further.

[6] https://www.icd10data.com/
[7] https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm
[8] https://www.who.int/standards/classifications/classification-of-diseases

4 Methodology and Networks

4.1 Network Notations

In academic literature, networks consist of two objects:

1. Nodes - refer to the entities
2. Edges- refer to the relationships between entities.

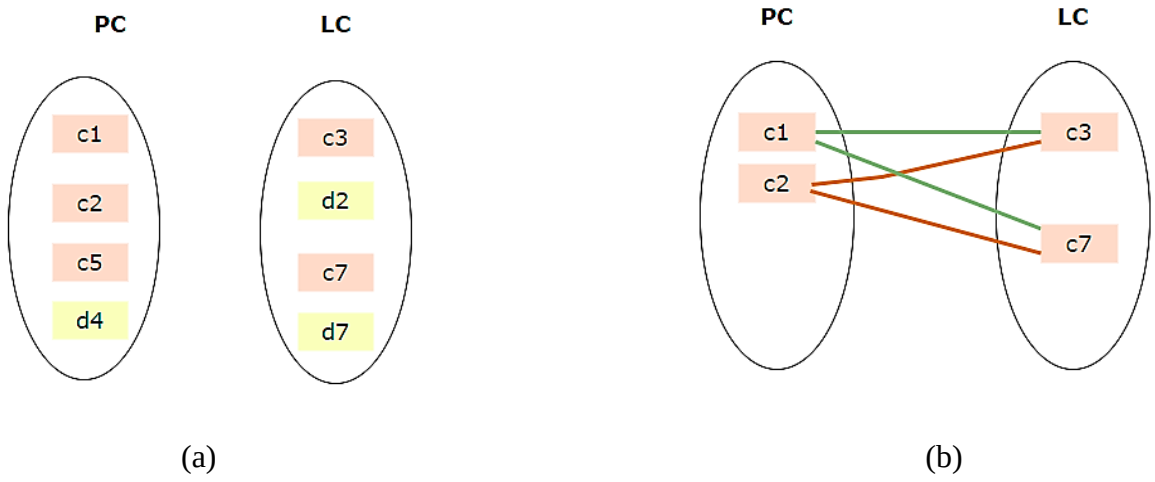
For instance, the node set N comprised elements of: $N = \{a, b, c, d \ldots .\}$ and the edge set would be a set of tuples $E = \{(a, b), (a, c), (c, d), \ldots\}$. There exist two types of graphs: directed and undirected. In the directed graphs the edges point in a direction, and in the undirected graphs the edges are bidirectional. When studying networks, edges are significant to evaluate. "*The heart of a graph lies in its edges, not in its nodes* "cited by John Quackenbush, Harvard School of Public Health. In our case, patient-diseases are a typical problem that can be designed by using the concept of graphs which leads us to networks. It is particularly important to point out that real-life scenarios are more complicated than just a representation of nodes and edges. In health care, the dynamic and variety of diseases needs more insights and thoughts to draw a graph for the diseases. For instance, there are some typical diseases/conditions that are considered chronic ones such as: asthma, diabetes, hypertension and so on. These chronic diseases cannot be caused by a particular condition like COVID-19. Chronic diseases are persistent and cannot be considered as long-COVID conditions. Meanwhile, there are other non-chronic diseases that might appear after a certain condition like COVID-19. These interpretations related to the environment we are studying are reflected in the way we design our graph or network. Therefore, let us define some notations used in our study:

Let P, represent unique patient $P = \{p_{1,}, p_{2,}, p_3, \ldots . p_n\}$, where $p_{1,}, p_{2,}\ldots$ represent different pediatric patients. Let C, represent the chronic set of conditions $C = \{c_{1,}, c_{2,}, c_3, \ldots . c_n\}$, where $c_{1,}, c_{2,}$ represent different chronic conditions. Let D, represent the set of non-Chronic

conditions $D = \{d_{1,}, d_{2,}, d_3, \ldots . d_n\}$, where $d_{1,}, d_{2,}$…represent different non-chronic conditions. In the next section, we will evaluate the connections among chronic diseases in two time periods pre-COVID and long COVID.

4.2 Chronic-DN: Chronic Diseases Network

As previously explained, chronic diseases are a specific case, and their presence is not necessarily caused by COVID-19. Therefore, we should treat them carefully when estimating any associations with COVID. For instance, in figure 4.1.a) we have a combination of chronic and non-chronic diseases in PC and LC. To determine the list of chronic diseases, we referred to the study [16].



(a) (b)

Figure 4.1: Chronic - Diseases Network. (a) All diseases in Pre-COVID and Long-COVID (b) Only chronic diseases in Pre-COVID and Long-COVID

The steps to build up our chronic diseases network are described below:

1. Create a master list that contains all patients and their chronic diseases in PC
2. Create a master list that contains all patients and their chronic diseases in LC
3. Create a master list that counts all the permutations of diseases in PC and LC for each of the pediatric patients. The master list will include PC, LC and count.

4. As a consequence of extracting permutations, we will have null values which are removed.
5. Finally, we apply a descending order by the frequency of the diseases that are associated with each other.

Figure 4.1.b) provides a simple chronic diseases network. The data modelling for this type of network is provided in appendix A.4. Table 4.1. shows the 10 most frequent sequences of chronic diseases in pre-COVID and long COVID.

Chronic – Diseases Associations for pediatric patients		
Pre COVID one year back in time (PC1)	**Long COVID (LC) from four weeks to one year**	**Frequency (n = 90680)**
J45- Asthma	J45- Asthma	6938(7.651%)
F41- Other anxiety disorders	F41 - Other anxiety disorders	2975(3.28%)
E66- Overweight and obesity	E66- Overweight and obesity	2476(2.73%)
F32 -Depressive episode	F32- Depressive episode	1121(1.236%)
J45- Asthma	E66 -Overweight and obesity	1068(1.178%)
F32-Depressive episode	F41- Other anxiety disorders	1039(1.146%)
F41- Other anxiety disorders	F32-Depressive episode	977(1.078%)
J45- Asthma	F41- Other anxiety disorders	951(1.049%)
H90- Conductive and sensorineural hearing loss	H90- Conductive and sensorineural hearing loss	898(0.99%)
E66-Overweight and obesity	J45- Asthma	896(0.988%)

Table 4.1: Top 10 most frequent sequences of chronic diseases in pre and long COVID

From these results we can observe that the most frequent chronic diseases appeared in pediatric patients in the timeframe pre and long COVID are asthma, anxiety, overweight and obesity followed by disorders of lipoprotein metabolism and other lipidemia's, anemias, primary hypertension, Type 1 diabetes, chronic kidney disease (CKD) and so on.

4.3 Complete – Diseases Network: Chronic- nonchronic diseases in pre-COVID and nonchronic diseases in long- COVID

This network is designed to evaluate non-chronic disease that might appear as long-term effects of COVID. As we previously mentioned, chronic diseases that have been diagnosed earlier are not of a main interest. We focus on addressing new conditions. Therefore, for each patient in long-COVID we extracted all chronic diseases and removed them in this master table. These records were used to update the table in pre-COVID. Basically, for each patient we are storing all chronic diseases in the time frame of pre-COVID. The figure 4.2 provides a visual representation of the heuristic applied. In Appendix A.5 is given the workflow used to implement this approach.

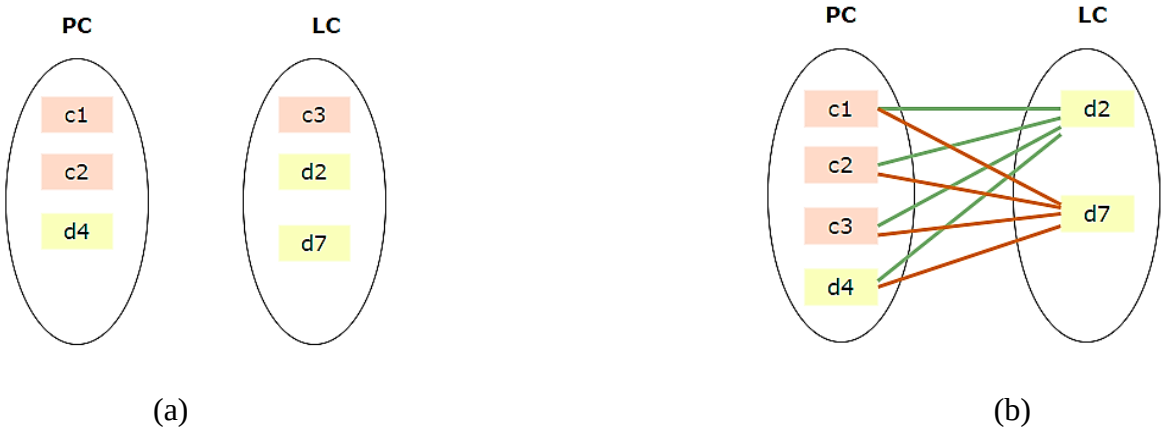


Figure 4.2: Complete- Disease Network. (a) All diseases in Pre-COVID and Long-COVID (b) Chronic-nonchronic diseases in PC(Pre-COVID) and nonchronic diseases in LC (Long-COVID)

In table 4.2, we have summarized the top 10 sequences of diseases generated for Complete - Diseases Network.

Complete- Diseases Network: nonchronic – chronic diseases in PC and nonchronic diseases in LC Network		
Pre COVID one year back in time (PC1)	**Long COVID (LC) from four weeks to one year**	**Frequency (n= 5424907)**
J06-Acute upper respiratory infections of multiple and unspecified sites	J06-Acute upper respiratory infections of multiple and unspecified sites	6841(0.126%)
R50-Fever of other and unknown origin	R50-Fever of other and unknown origin	6479(0.119%)
R50-Fever of other and unknown origin	J06-Acute upper respiratory infections of multiple and unspecified sites	5972(0.11%)
J02-Acute pharyngitis	J02-Acute pharyngitis	5345(0.098%)
F90-Attention-deficit hyperactivity disorders	F90-Attention-deficit hyperactivity disorders	5088(0.094%)
R05-Cough	J06-Acute upper respiratory infections of multiple and unspecified sites	4732(0.0872%)
R05-Cough	R05-Cough	4728(0.0871%)
R50-Fever of other and unknown origin	R05-Cough	4667(0.086%)
J06-Acute upper respiratory infections of multiple and unspecified sites	R50-Fever of other and unknown origin	4194(0.077%)
R05-Cough	R50-Fever of other and unknown origin	4119(0.076%)

Table 4.1: Top 10 most frequent sequences of diseases for Complete – Diseases Network in pre and long COVID

Previously we observed that positive cases in PC and LC are equally distributed among females and males. Therefore, we considered age as a parameter to split the Complete – Diseases Network and estimate the differences between them. According to the CDC[9] , they group all children and teens in three main categories to provide guidance for the vaccination authorization. Therefore, we used this reference to evaluate three age categories:

a. Under 4 years old

[9] https://www.cdc.gov/coronavirus/2019-ncov/vaccines/recommendations/children-teens.html

b. 5 to 11 years old
c. 12 to 17 years old

It is very crucial to point out that for patients under 4 years old is very difficult to capture specific symptoms such as loss of smell, taste, or any other that cannot be reported by the patient. On the other hand, among the two other groups the body anatomy: organs, structure, systems, and psychology are very different. This has a direct impact on the potential diseases that might be appear. We will show that some patients at age 15,16,17 appeared to be experiencing conditions/symptoms related to the pregnancy. In appendix C, we have summarized the results for these three categories. In addition, we have visualized the top 50 sequences of diseases using a Sankey diagram to present this information (Appendix D).

Some observations:

1. For the category under 4 years old, the most frequent diseases appeared in long COVID were: fever, acute upper respiratory infections, suppurative and unspecified Otis media, symptoms and signs involving the circulatory and respiratory system, viral infections, nausea and vomiting, abnormalities of breathing, symptoms and signs involving digestive system and abdomen.
2. For the category 5 to 11 years old, the most frequent diseases appeared in long COVID were: acute upper respiratory infections, acute pharyngitis, cough, vasomotor and allergic rhinitis, suppurative and unspecified Otis media, functional intestinal disorders.
3. For the category 12 to 17 years old, the most frequent diseases appeared in long COVID were: acute pharyngitis, deficit- hyperactivity disorders, abdominal and pelvic pain, vasomotor and allergic rhinitis, other joint disorder, reaction to severe stress, sleep disorders.

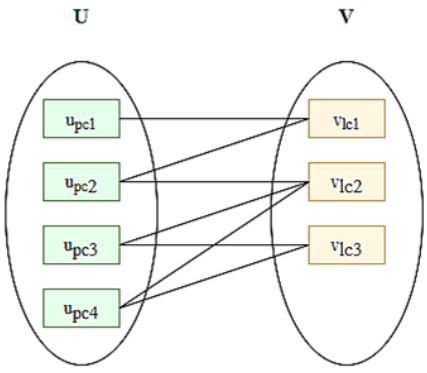
4.4 Bipartite graphs and projections

Bipartite graph or biograph is a network where nodes can be divided into two sets U and V, and a list of edges E. In table 4.2, we summarized some notations related to our graphs.

Symbols	Meaning
$G(U_{PC}, V_{LC}, E)$	Pre COVID-Diseases -Long COVID Diseases graph
U_{PC}	Set of diseases in Pre COVID $\{u_{PC1}, u_{PC2}, u_{PC3}\}$
V_{LC}	Set of diseases in Long COVID$\{v_{LC1}, v_{LC2}, v_{LC3}\}$
n_1	Number of diseases in Pre-COVID (PC1)
n_2	Number of diseases in Long-COVID (PC1)
B	Bi- adjacency matrix of G

Table 4.2: Symbols and Notations

Complete – Diseases Network can be modeled as a bipartite graph $G(U_{PC1}, V_{LC}, E)$, where U-the set of diseases in Pre-COVID and V-the set of diseases in Long-COVID. For $u_m \epsilon U$, $v_n \epsilon V$, and $u_m v_n \epsilon E(G)$, if there exists an association among diseases in pre and long COVID. In figure 4.3. we can see a bipartite graph, where:



$$B = \begin{array}{c} \\ u_{pc1} \\ u_{pc2} \\ u_{pc3} \\ u_{pc4} \end{array} \begin{array}{c} \begin{array}{ccc} v_{lc1} & v_{lc2} & v_{lc3} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

Figure 4.3: Bipartite graph representing the diseases in pre-COVID and long-COVID and its bi-adjacency matrix

$U_{PC} = \{u_{PC1}, u_{PC2}, u_{PC3} \ldots\}$ and $V_{LC=} = \{v_{LC1}, v_{LC2}, v_{LC3} \ldots\ldots\}$ and its bi-adjacency matrix.

Bipartite graphs are difficult to analyze due to lack of techniques, hence it is converted into a unipartite graph or known as projections. To generate projections (unipartite graphs), we will use its bi-adjacency matrix. This matrix is filled out with 1-s if there is any connection among diseases in pre- COVID and long COVID. Otherwise, it is filled out with 0-s.

5 Community Detections

As explained in chapter 2.3., the importance of community detection algorithms as a key component in understanding and analyzing the structure of complex networks to extract meaningful information. The disease projections we generated from the two-time frames: pre-COVID and Long COVID are very dense to obtain and evaluate association and similarity among diseases. In table 5.1, we have summarized the content of these graphs for the Complete-Diseases Network. As we can observe the number of edges and average degree are exceptionally large to interpret and visualize a graph. Therefore, we have applied community detection algorithms appropriate to our problem.

Complete - Diseases Network	**Nodes**	**Edges**	**Average degree**
Projections in Pre COVID-19	253	27,136	214.514
Projections in Long COVID-19	1328	849,470	1279.322

Table 5.1: Summary of graph projections in pre and long COVID

In figure 5.1, we have visualized the projections in long COVID, but it is evident that is impossible to extract pragmatic outcomes. This led us to community detection algorithms.

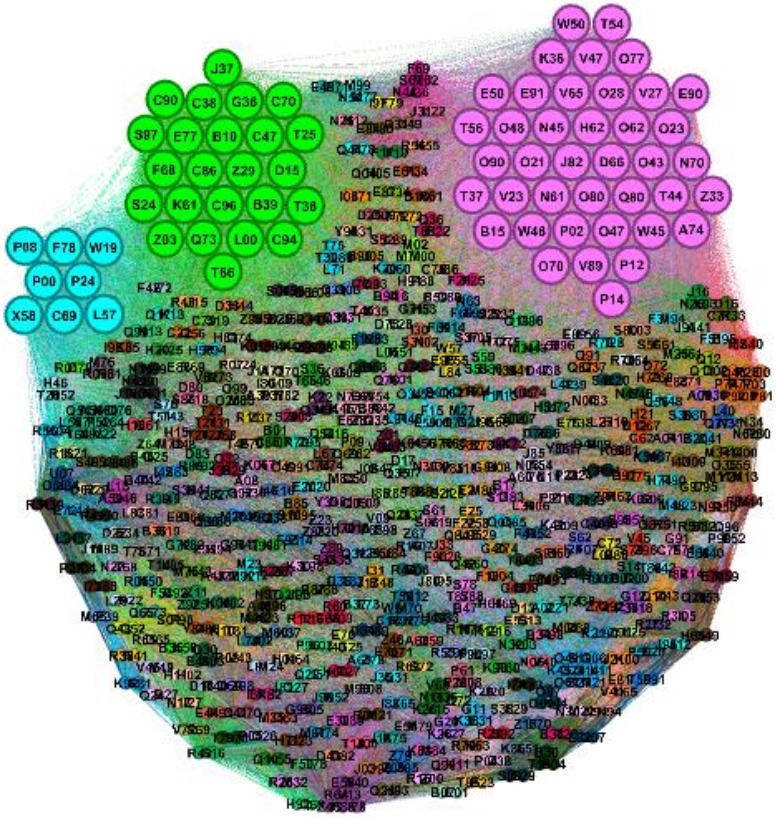


Figure 5.1: Projections in long COVID by filtering out the edges above the average degree (avg. degree = 1279)

5.1 Louvain and Leiden algorithms

To measure the structure of a graph and the density of connections between nodes in one community we use the modularity score. A high value of modularity score is interpreted as well-connected nodes within a community and only few connections directed outwards. The Louvain algorithm [17] is a simple, intuitive, and efficient algorithm used to identify communities in large and complex networks. This algorithm is designed to maximize the modularity score. The Louvain algorithm has been used in many studies [18], [19], [20], [21], [22]. However, a general problem observed in this algorithm is that it can produce clusters that are badly connected and that should have been split up into multiple clusters. To overcome this issue, Leiden algorithm [23] guarantees that clusters are well-connected. Many studies have

analyzed and compared these algorithms [24]. They have proven that the Leiden algorithm is more stable and time-efficient in detecting communities in large networks. We applied both algorithms in our projections and the results are summarized in table 5.2:

Complete Diseases Network	Louvain (Modularity Class) Resolution= 0.2			Leiden algorithm Constant Potts Model (CPM)		
	Nodes	Edges	Communities	Nodes	Edges	Communities
PC1-Projections	77 (30.43%)	1555 (5.73%)	76	251 (99.21%)	268857 (98.97%)	3
LC- Projections	295 (22.21%)	33155 (3.9%)	402	1185 (89.23%)	700,867 (82.51)	8

Table 5.2: The results from Louvain and Leiden algorithms

5.2 Results

We have visualized the top 5 clusters generated by Louvain algorithm in figure 5.2 for pre-COVID and figure 5.3 for long-COVID. The nodes stand for the 3-digit part from ICD-10 code.

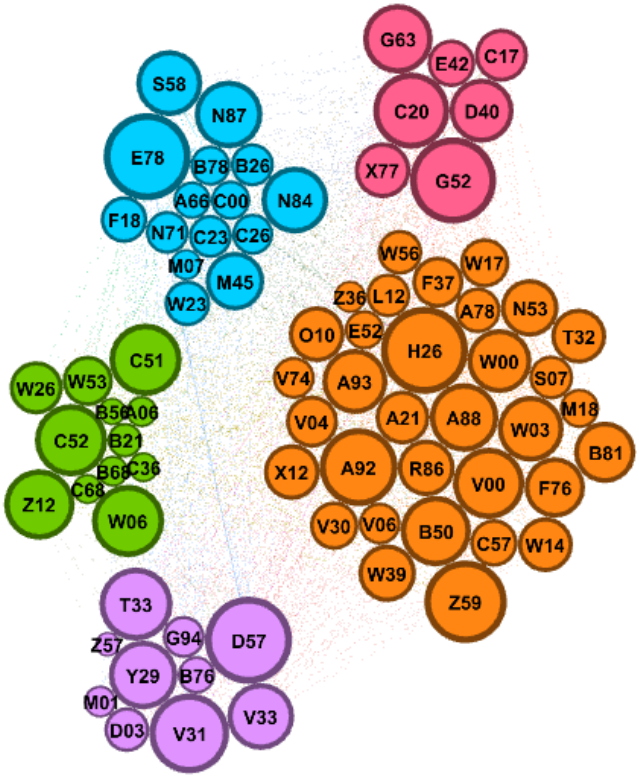


Figure 5.2: Top 5 communities generated by Louvain algorithm in pre-COVID

In the 1-st community, figure 5.3., which make up to 13.04% we observed the dominant conditions as below:

Figure 5.3: 1-st community using Louvain algorithm in pre-COVID

- Cataract- cloudy or blurry vision. Some reasons why a child may develop distortion of vision are: genetic fault (Down's syndrome), some infections picked up by the mother during pregnancy (rubella or chickenpox).
- Plasmodium Falciparum and Q fever– the symptoms related to it are fever, chills, sweats, headaches, nausea, body aches and it can cause severe malaria and aggravate up to anemia.
- Pre-existing essential hypertension
- Viral infectious of central nervous system
- Tularemia is a rare infectious disease. Symptoms may vary from *fever, chills, and swollen lymph nodes up to skin or mouth ulcers, diarrhea, muscle aches, joint pain, cough, and weakness*. This can be caused by mosquitos, and we found other conditions such as: A.92: mosquito born viral fevers, and A.93: arthropod-borne viral fevers e

present in this community. These conditions are very typical in the tropical areas, which is an indication how the location can have a high impact.

- Intestinal helminthiasis which causes *abdominal pain, bloody diarrhea, nausea, vomiting, headache* and so on.

In the 2-nd community, figure 5.4, which composes 5.93% of the network, we observed the dominant conditions as below:

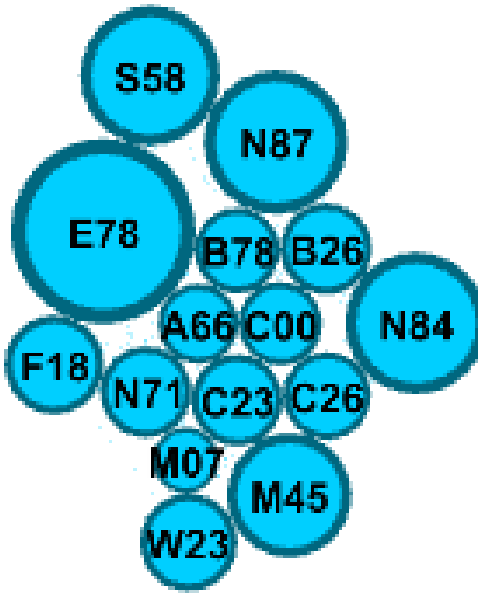


Figure 5.4: 2-nd community using Louvain algorithm in pre-COVID

- disorders of lipoprotein metabolism and other lipidemia's- not having enough enzymes to break down lipids. Or the enzymes may not work properly, and your body can't convert the fats into energy. They cause a harmful amount of lipids to build up in your body. Over time, that can damage your cells and tissues, especially in the brain, peripheral nervous system, liver, spleen, and bone marrow [25] Obesity, the metabolic syndrome and diabetes are commonly associated with disorders of lipid and lipoprotein metabolism.[26]

- AS (ankylosing spondylitis) which appears in form of lower back pain, hip pain, neck pain, difficulty breathing, fatigue, loss of appetite, abdominal pain and diarrhea. In addition, we observed Enteropathic Arthritis (EA) which is classified as one type of Spondylitis
- Inhalant related disorders- used to code substance dependency. [10]
- Malignant neoplasm of lip/gallbladder and other ill-defined digestive organs
- Mumps
- Yaws – is a chronic skin infection that also may infect bones in its late stages. It mostly affects children in tropical regions.

In the 3-rd community, figure 5.5, which makes up to 4.74% of the network, we observed the dominant conditions as below:

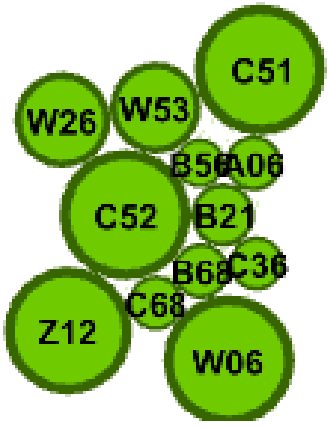


Figure 5.5: 3-rd community using Louvain algorithm in pre-COVID

- Malignant of neoplasm/other unspecified urinary organs
- Mumps
- Chronic embolism and thrombosis of right subclavian vein(SCVT) – is a potentially life-threatening pathology if not treated in a timely manner[27]

[10] https://icd.codes/icd10cm/F18

- Common Intestinal Parasites: Amebiasis and Taeniasis – very typical in the tropical areas in school age children.

In the 4-th community, figure 5.6, which contains 3.95% of the network, we observed the dominant conditions as below:

Figure 5.6: 4-th community using Louvain algorithm in pre-COVID

- Sickle cell disorders- is related to the shape of the red blood cells which carry the oxygen to all parts of the body. Some of the symptoms refer to infections, pain and fatigue. Melanoma in situ is another disease of cancer cells in the top layer of the skin.
- Frostbite- slight changes in skin color. This is very common on the fingers, toes, nose, ears, checks, and chin. A similar finding was reported in an article in National Geographic [28]
- Intestinal parasites like Hookworm.

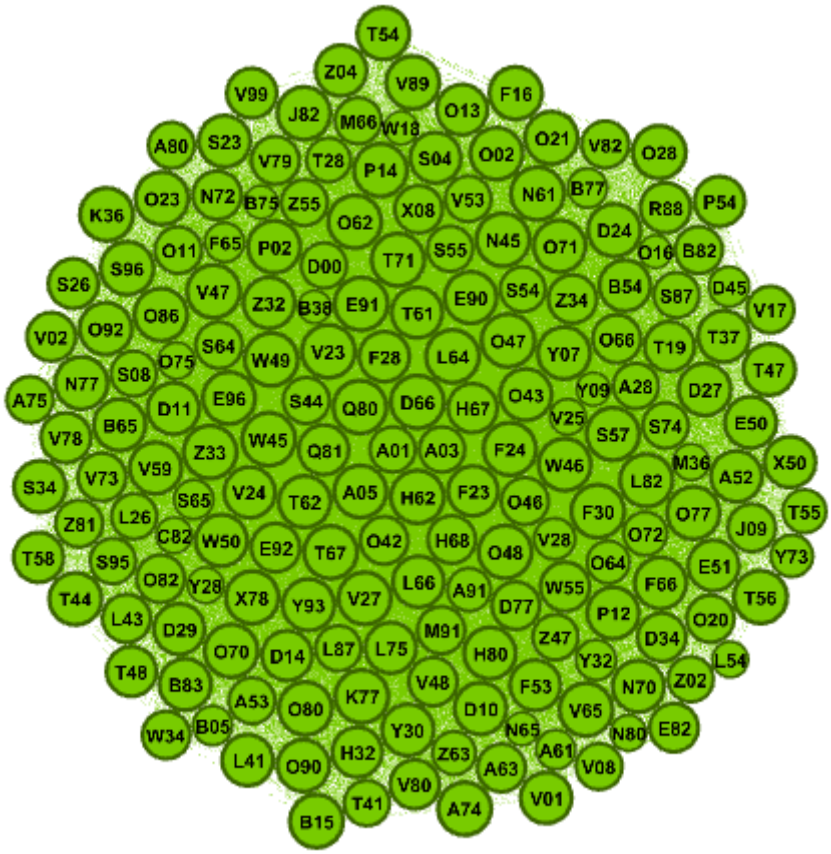
In the 5-th community, figure 5.7, which contains 2.77% of the network, we observed the dominant conditions as below:

Figure 5.7: 5-th community using Louvain algorithm in pre-COVID

- Disorders of cranial nerves- which might cause symptoms that include intense pain, vertigo, hearing loss, weakness, or paralysis. These nerves are very delicate in the human nervous system.[29]
- Malignant neoplasm of rectum/ small intestine
- Marasmic kwashiorkor- is caused by protein deficiency. The typical symptoms are: weight loss, dry skin and eyes, brittle hair, diarrhea, lower immunity and so on.

The communities generated by the Louvain algorithm in long COVID are provided in Figure 5.8. We grouped the conditions by the category, and the results are summarized in Table 5.3. However, some observations for each community:

Figure 5.8: Top 5 communities generated by Louvain algorithm in long COVID (a) 1-st community (b) 2-nd community (c) 3-rd community (d)4-th community (e) 5-th community



(a) 1-st community

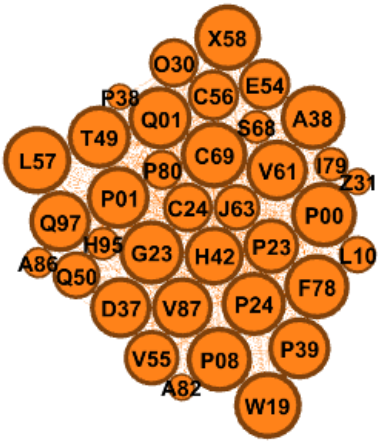
1. In the 1-st community which contains 14.01% of this network, Figure 5.8. a), we noticed:

- The range of V00-Y99- external causes of morbidity include *20.10%* in this cluster.
- We observed many nodes(conditions) related to the pregnancy. All the codes in the range (O10-O9A). We can list some of them: O72, O16, O21, O11, O13, O86 etc. These conditions comprise *14%* of this community.
- We didn't find out any disease related to the nervous system(G00-G99) or circulatory system (I00-I99)
- We found out that infectious and parasitic diseases (A00-B99) made up *11.11%*.
- Mental and behavioral disorders involve *4.76%* of the nodes in this cluster.

2. In the 2-nd community, we noticed that the category of diseases with the highest the percentage of 24.64% is neoplasm. This category is related to the cancer. It is followed by infectious and parasitic diseases ranking with 17.39%.

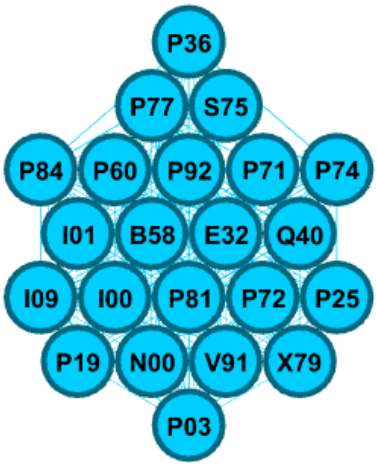
(b) 2-nd community

3. In the 3-rd community, the presence of conditions originating in the perinatal period is *22.85%*. Even in this community, the conditions related to the cancer (neoplasm) make up *11.42%*.



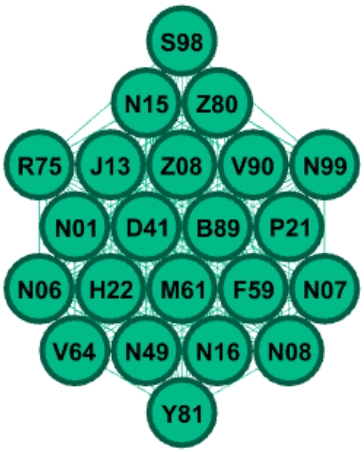
(c) 3-rd community

4. In the 4-th community, the dominant group is related to the conditions originating from the perinatal period with a percentage of *56.52%*.



(c) 4-th community

5. In the 5-th community, we noticed diverse conditions with a dominance in the category in the genitourinary system with a percentage of *38.1%*.



(d) 5-th community

ICD-10 category	Communities using Louvain Algorithm in LC						Leiden in PC
	1-st	2-nd	3-rd	4-th	5-th	Top 5 communities	1-st community
A00-B99 Certain infectious and parasitic diseases	21	12	3	1	1	38	3
C00-D49 Neoplasms	10	17	4	None	1	32	3
D50-D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	2	2	None	None	None	4	11
E00-E89 Endocrine, nutritional, and metabolic diseases	2	4	1	1	None	8	13
F01-F99 Mental, Behavioral and Neurodevelopmental disorders	9	1	1	None	1	12	8
G00-G99 Diseases of the nervous system	None	2	1	None	None	3	14
H00-H59 Diseases of the eye and adnexa	1	None	1	None	1	3	12
H60-H95 Diseases of the ear and mastoid process	5	None	1	None	None	6	3
I00-I99 Diseases of the circulatory system	None	None	1	3	None	4	26
J00-J99 Diseases of the respiratory system	2	2	1	None	1	6	8
K00-K95 Diseases of the digestive system	2	3	None	None	None	5	5
L00-L99 Diseases of the skin and subcutaneous tissue	9	3	2	None	None	14	None
M00-M99 Diseases of the musculoskeletal system and connective tissue	3	1	None	None	1	5	18
N00-N99 Diseases of the genitourinary system	7	None	None	1	8	16	12
O00-O9A Pregnancy, childbirth, and the puerperium	26	1	1	None	None	28	3
P00-P96 Certain conditions originating in the perinatal period	4	2	8	13	None	27	4
Q00-Q99 Congenital malformations, deformations, and chromosomal abnormalities	2	1	3	1	None	7	None
R00-R99 Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	1	2	None	None	1	4	2
S00-T88 Injury, poisoning, and certain other consequences of external causes	31	6	1	1	1	40	2
V00-Y99 External causes of morbidity	38	5	5	1	3	52	5
Z00-Z99 Factors influencing health status and contact with health services	9	5	1	1	2	18	3

Table 5.3: Summarize the conditions by category obtained from Long-COVID Projections

As we previously explained, the Leiden algorithm is used to compare the results with the Louvain algorithm. The parameters used to generate the graph using Leiden algorithm in pre COVID are summarized in Figure 5.9. We have analyzed the 1-st community of 99.21%.

Configuration

Algorithm	Leiden
Quality Function	Constant Potts Model (CPM)
Resolution	0.5
Number of iterations	10
Number of restarts	1
Random seed	0

Results

Quality	0.9968910922986691
Number of clusters	3

a)

Configuration

Algorithm	Leiden
Quality Function	Constant Potts Model (CPM)
Resolution	0.5
Number of iterations	10
Number of restarts	1
Random seed	0

Results

Quality	0.9887117944611546
Number of clusters	4

b)

Figure 5.9: The parameters used to run Leiden algorithm a) pre-COVID b) long-COVID

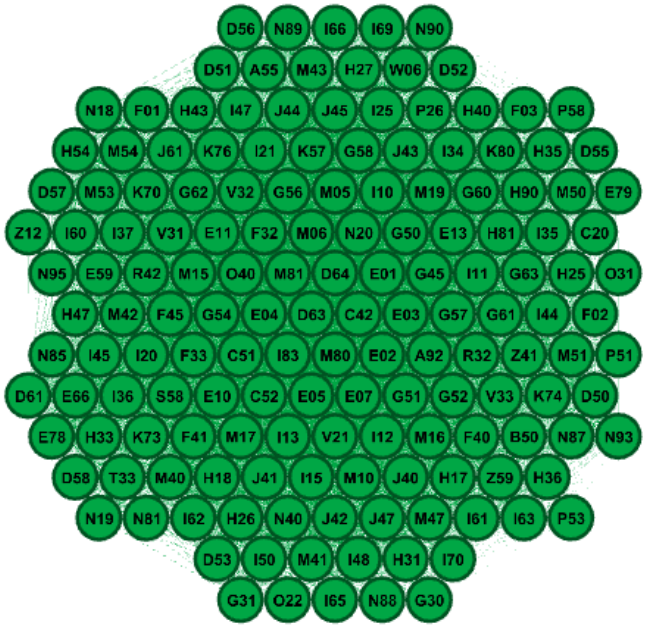


Figure 5.10: 1-st community generated by Leiden algorithm in pre-COVID filtering out by average degree (avg. degree = 214)

In this community, we observed the presence of some chronic diseases which are very critical like: CKD (chronic kidney disease), Essentially(primary) hypertension, secondary hypertension, type 1 diabetes, and type 2 diabetes. These are some conditions we didn't observe by using the Louvain algorithm in the pre-COVID. On the other hand, we observed these conditions from the results in the Chronic Diseases Network. This might be related to the fact that Leiden algorithm guarantees the connectedness of the nodes. Among the dominant categories we can list diseases of circulatory system (16.77%), diseases of musculoskeletal system and connective tissue (11.61%), nervous system (9%) and other diseases.

The graph obtained from the Leiden algorithm in long COVID was very dense, but we could capture the diseases(nodes) with highest degree. We observed that 25% were conditions related to pregnancy and diseases originating in the perinatal period, 13.24% neoplasm and 7.35% infectious and parasitic diseases.

6 Limitations and Future Research

6.1 Limitations

Healthcare is a very dynamic, challenging, and complex environment to study. Dealing with big data environments, and privacy policies on accessing/extracting any information outside the environment restricts and creates some challenges to overcome. This is a necessary trade-off to conduct a study based on real data.

In this study some challenges and restrictions we had to manage:

1. Apply some heuristic to reduce the search-space and dimensions of the problem.
2. The environment is extremely strict on privacy policies. This limited our work to focusing only on disease network rather than considering any patient association. In addition, we had to do workarounds in many cases to test our results. The database does not allow to view more than 1000 records unless you download any table.
3. Computational power was 8GB of RAM, which is not sufficient to work in an environment with big data. It becomes slow, unresponsive and causes delays.

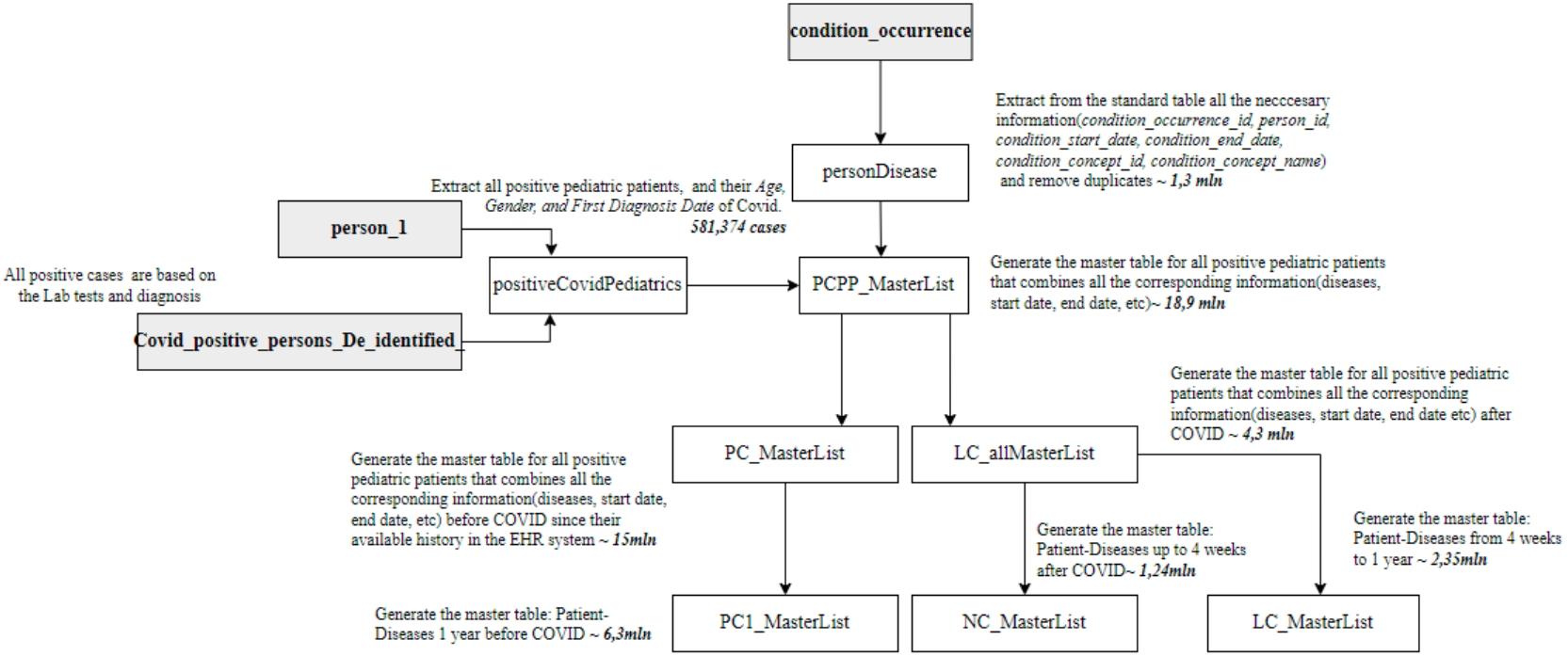
6.2 Future research

N3C environment is comprehensive and rich with data and collects records from multiple sites. Subsets can be studied such as diseases related to autoimmune systems, chronic diseases like diabetes, hypertension etc. Another study can be conducted to evaluate the impact of vaccines in the pediatric domain. We know that patients under 18 years old started to get vaccinated late. Therefore, we need to have more records before conducting any study.

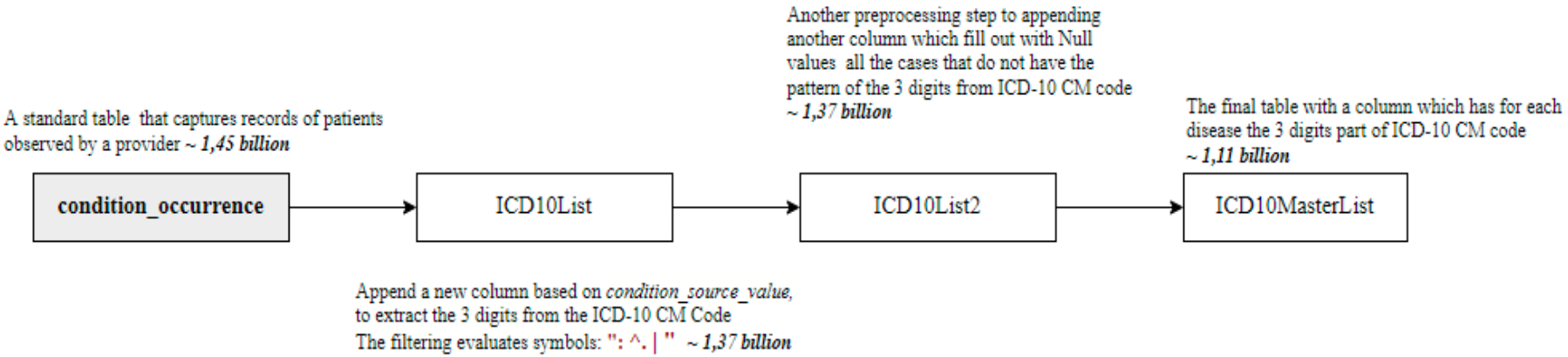
Appendices

Appendix A: Workflow

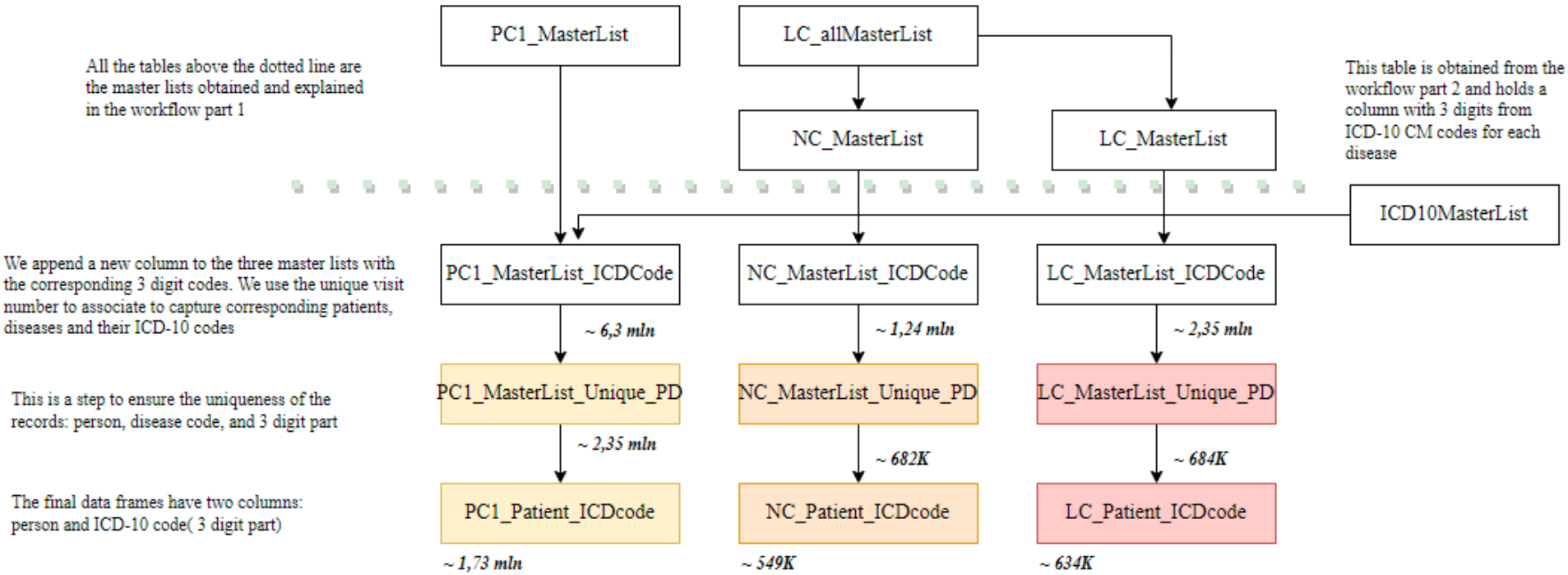
A.1: Generating master table (Patient- Diseases) in three timeframes



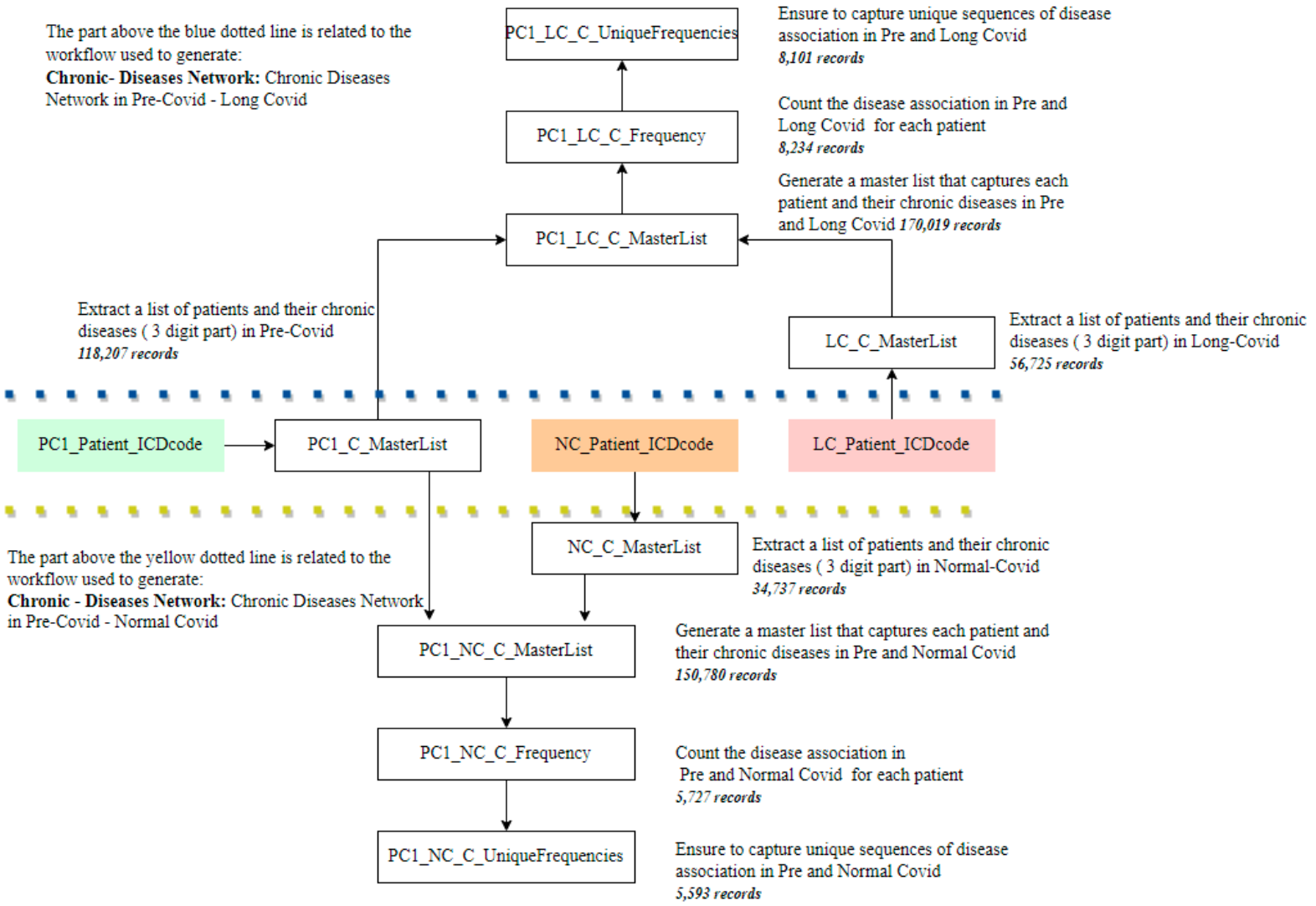
Appendix A.2: Extracting 3 digits from ICD-10 CM codes



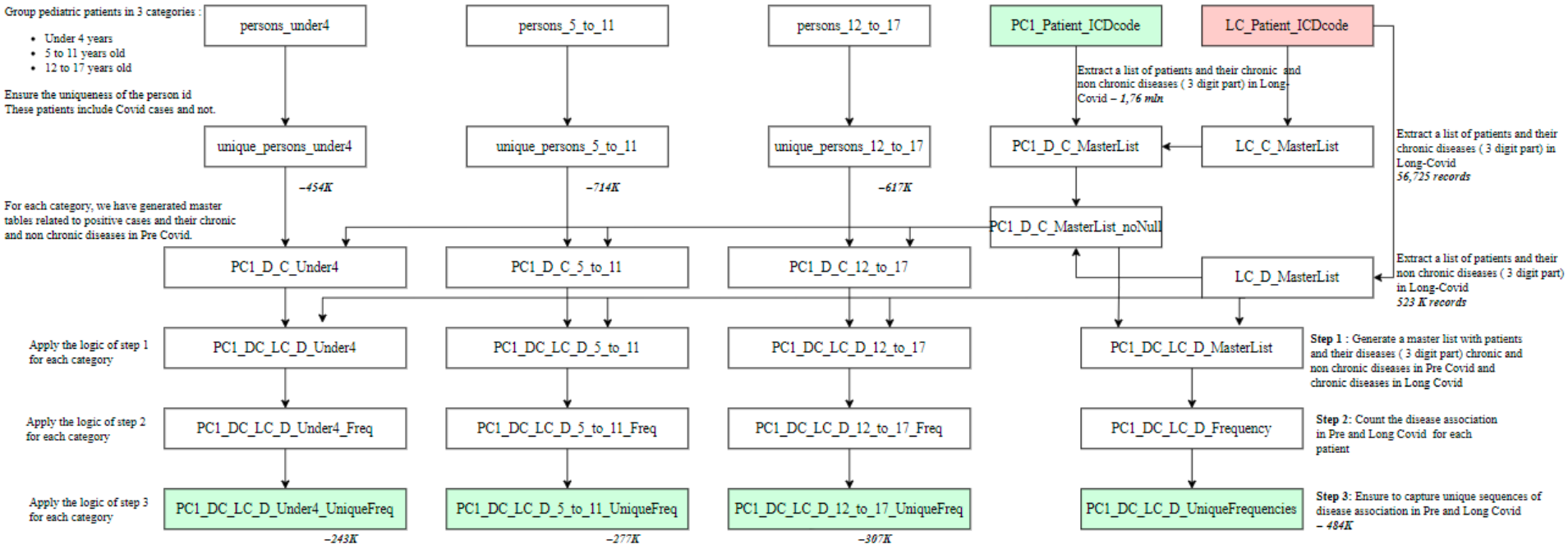
Appendix A.3: Generating three master lists with ICD-10 CM codes (3-digit part)



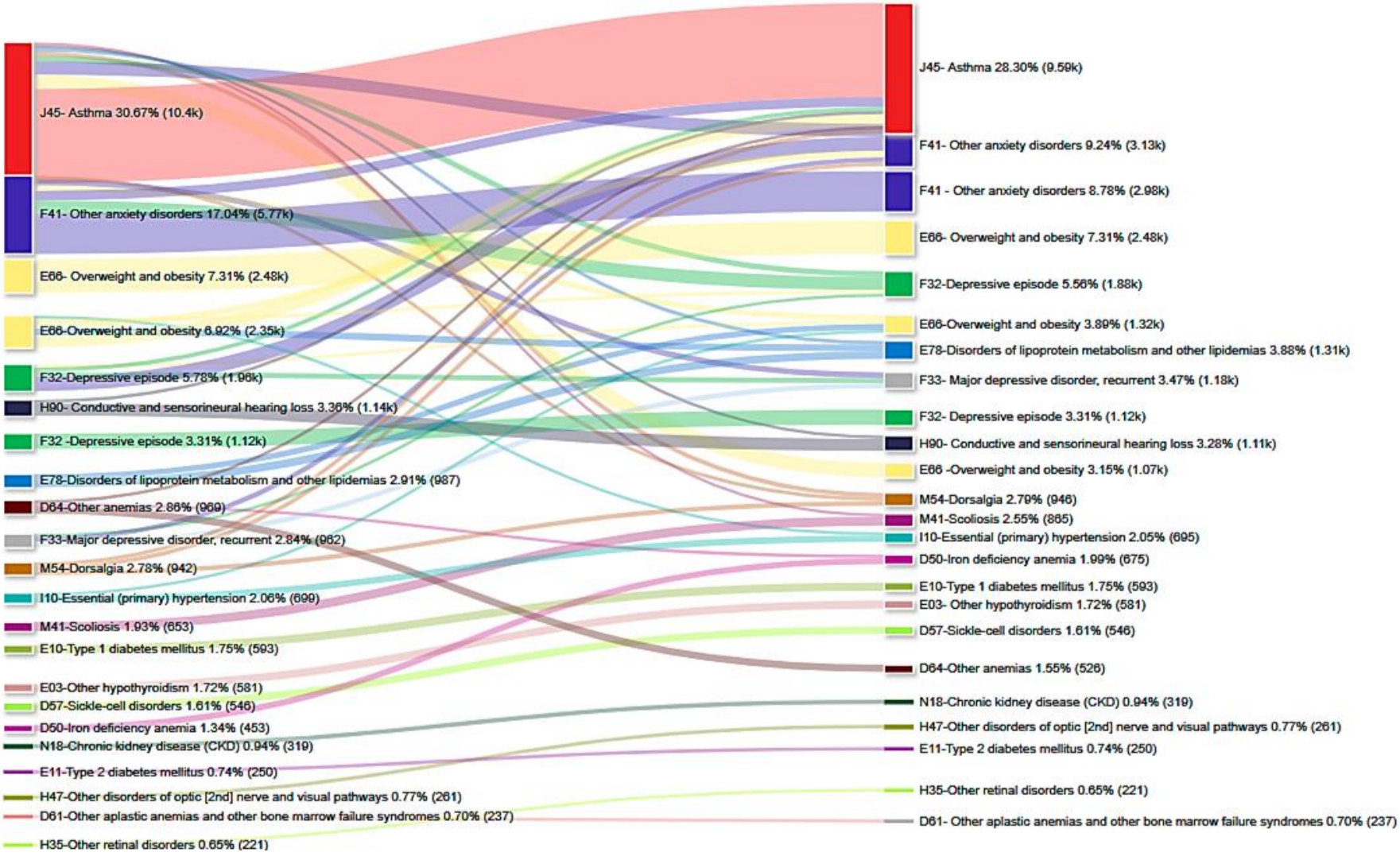
Appendix A.4: Chronic – Diseases Network



Appendix A.5: Complete – Diseases Network



Appendix B: Chronic -Diseases Network



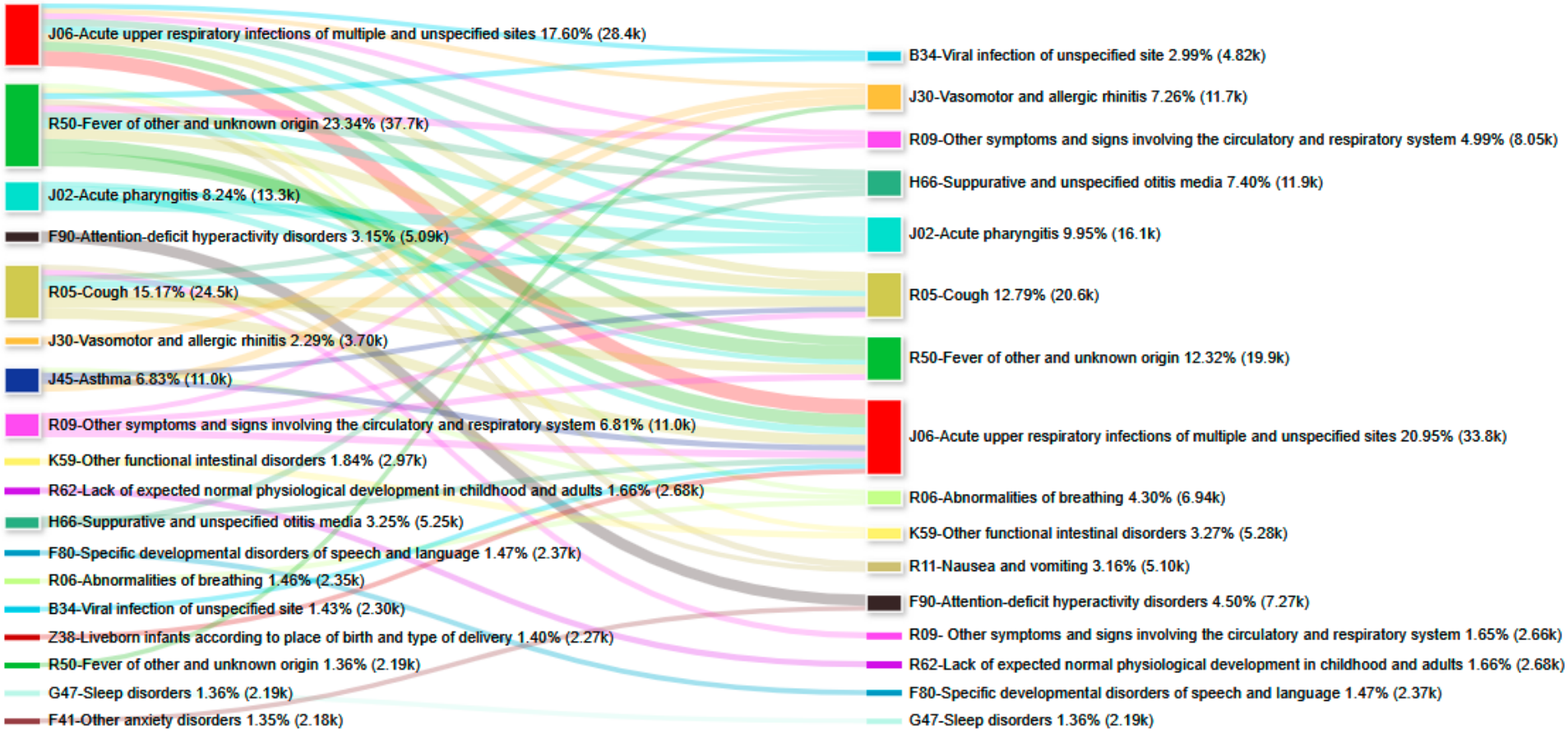
Appendix C: Complete- Diseases Network for age categories

Complete – Diseases Network AGE – Under 4 years old		
Pre-COVID one year back in time (PC1)	**Long COVID (LC) from four weeks to one year**	**Frequency (n= 1,888,041)**
R50-Fever of other and unknown origin	R50-Fever of other and unknown origin	3791(0.201%)
R50-Fever of other and unknown origin	J06-Acute upper respiratory infections of multiple and unspecified sites	3393(0.1797%)
J06-Acute upper respiratory infections of multiple and unspecified sites	J06-Acute upper respiratory infections of multiple and unspecified sites	3329(0.1763%)
J06-Acute upper respiratory infections of multiple and unspecified sites	R50-Fever of other and unknown origin	2621(0.1388%)
R50-Fever of other and unknown origin	R05-Cough	2448(0.1297%)
R50-Fever of other and unknown origin	H66-Suppurative and unspecified otitis media	2419(0.1281%)
R05-Cough	J06-Acute upper respiratory infections of multiple and unspecified sites	2400(0.1271%)
R05-Cough	R50-Fever of other and unknown origin	2285(0.121%)
Z38-Liveborn infants according to place of birth and type of delivery	J06-Acute upper respiratory infections of multiple and unspecified sites	2256(0.1195%)
J06-Acute upper respiratory infections of multiple and unspecified sites	H66-Suppurative and unspecified otitis media	2203(0.1167%)

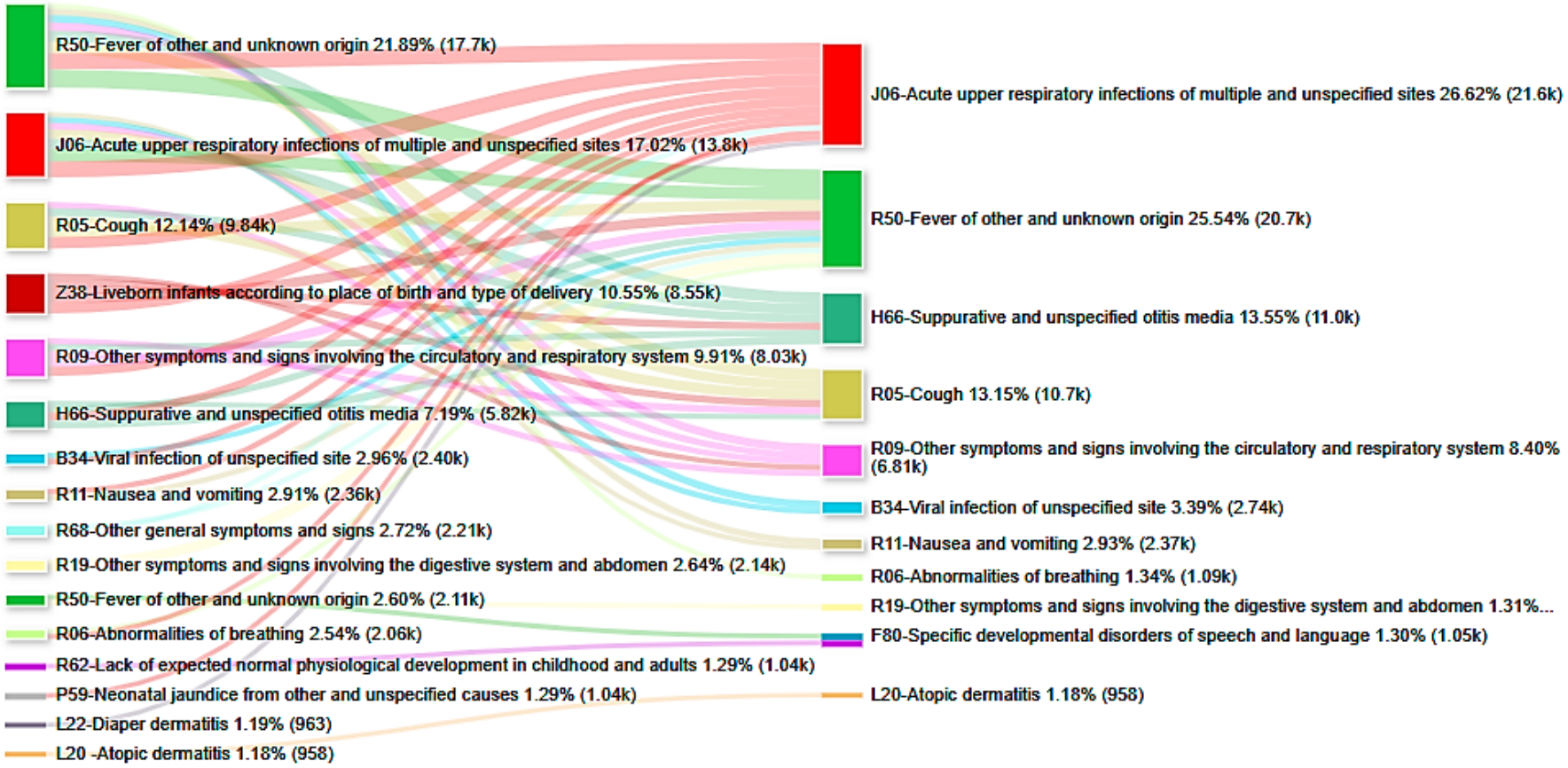
Complete – Diseases Network: AGE – 5 to 11 years old		
Pre-COVID one year back in time (PC1)	**Long COVID (LC) from four weeks to one year**	**Frequency (n=1,708,469)**
J02-Acute pharyngitis	J02-Acute pharyngitis	2228(0.1304%)
J06-Acute upper respiratory infections of multiple and unspecified sites	J06-Acute upper respiratory infections of multiple and unspecified sites	2173(0.1272%)
R50-Fever of other and unknown origin	R50-Fever of other and unknown origin	1971(0.1154%)
R50-Fever of other and unknown origin	J02-Acute pharyngitis	1913(0.112%)
R50-Fever of other and unknown origin	J06-Acute upper respiratory infections of multiple and unspecified sites	1863(0.109%)
F90-Attention-deficit hyperactivity disorders	F90-Attention-deficit hyperactivity disorders	1745(0.1021%)
R05-Cough	R05-Cough	1689(0.0987%)
J30-Vasomotor and allergic rhinitis	J30-Vasomotor and allergic rhinitis	1648(0.0965%)
R50-Fever of other and unknown origin	R05-Cough	1603(0.0938%)
J45- Asthma	J30-Vasomotor and allergic rhinitis	1593(0.0932%)

Complete – Diseases Network: AGE – 12 to 17 years old		
Pre Covid one year back in time (PC1)	**Long Covid (LC) from four weeks to one year**	**Frequency (n= 1,828,397)**
F90-Attention-deficit hyperactivity disorders	F90-Attention-deficit hyperactivity disorders	3337(0.1825%)
J02-Acute pharyngitis	J02-Acute pharyngitis	2901(0.1587%)
J30-Vasomotor and allergic rhinitis	J30-Vasomotor and allergic rhinitis	1819(0.0995%)
J45- Asthma	J30-Vasomotor and allergic rhinitis	1737(0.095%)
F41- Other anxiety disorders	F90-Attention-deficit hyperactivity disorders	1608(0.0879%)
R05-Cough	J02-Acute pharyngitis	1433(0.0784%)
J06-Acute upper respiratory infections of multiple and unspecified sites	J02-Acute pharyngitis	1377(0.0753%)
J02-Acute pharyngitis	J06-Acute upper respiratory infections of multiple and unspecified sites	1365(0.0747%)
J06-Acute upper respiratory infections of multiple and unspecified sites	J06-Acute upper respiratory infections of multiple and unspecified sites	1339(0.0732%)
R50-Fever of other and unknown origin	J02-Acute pharyngitis	1312(0.0718%)

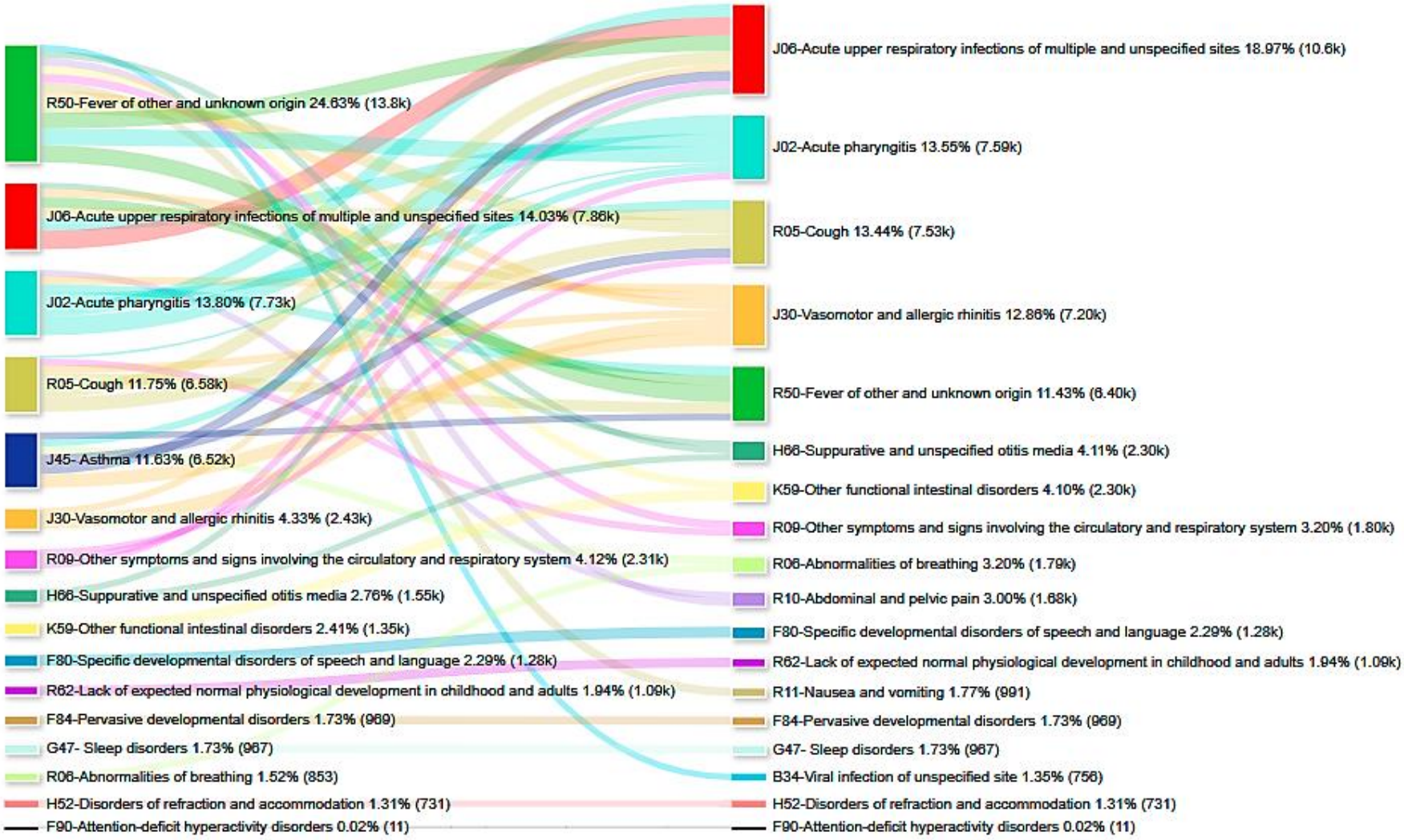
Appendix D: Complete- Diseases Network Sankey Diagrams



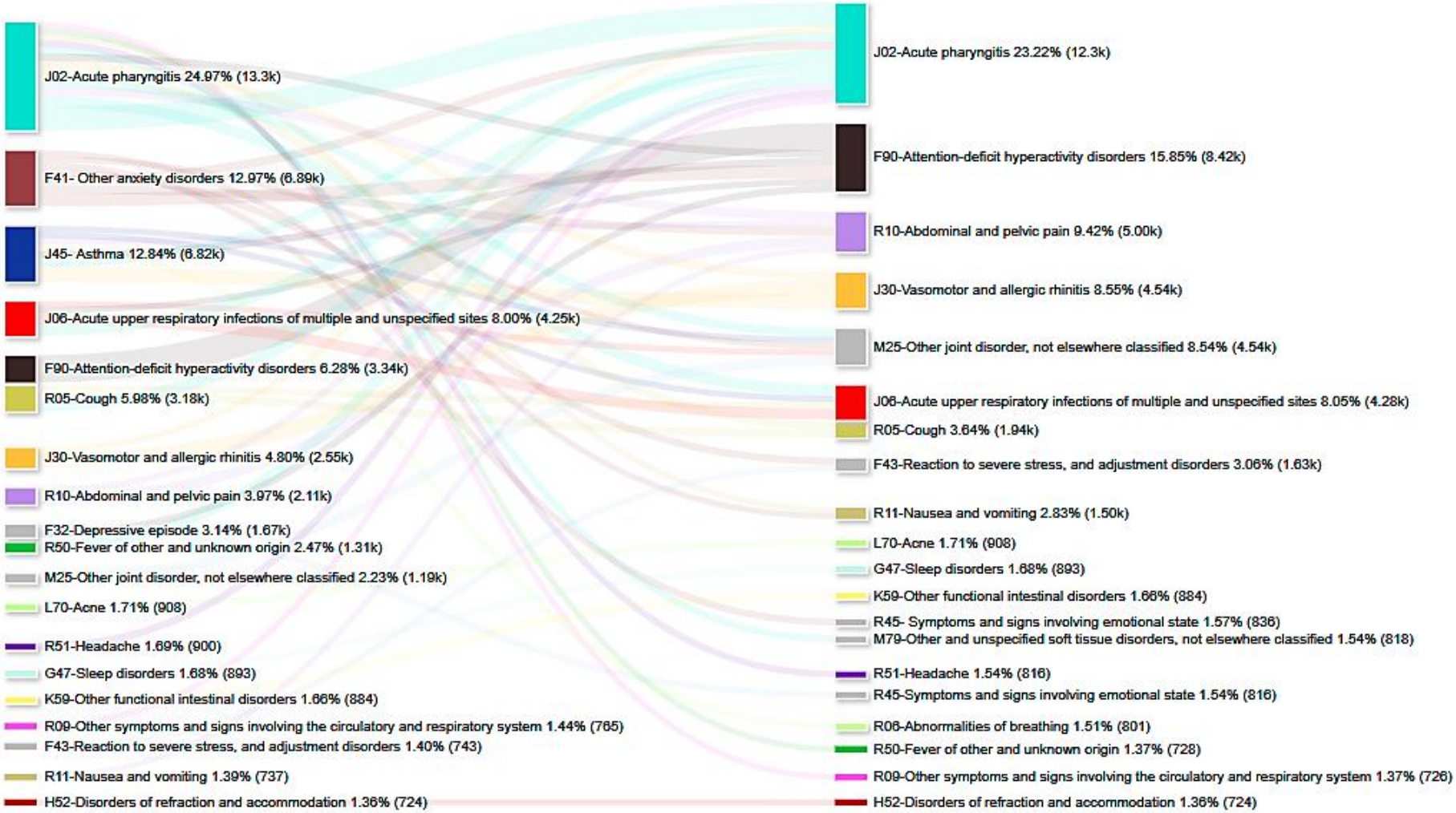
Age Category: Under 4 years



Age Category: 5 to 11 years



Age category: 12 to 17 years



References

[1] O. Irfan, F. Muttalib, K. Tang, L. Jiang, Z. S. Lassi, and Z. Bhutta, “Clinical characteristics, treatment and outcomes of paediatric COVID-19: a systematic review and meta-analysis,” *Arch Dis Child*, vol. 106, no. 5, p. 440, May 2021, doi: 10.1136/archdischild-2020-321385.

[2] A. Göktuğ *et al.*, “Evaluation of Epidemiological, Demographic, Clinical Characteristics and Laboratory Findings of COVID-19 in the Pediatric Emergency Department,” *Journal of Tropical Pediatrics*, vol. 67, no. 4, p. fmab066, Aug. 2021, doi: 10.1093/tropej/fmab066.

[3] S. Chahar and P. K. Roy, “COVID-19: A Comprehensive Review of Learning Models.,” *Arch Comput Methods Eng*, pp. 1–26, Sep. 2021, doi: 10.1007/s11831-021-09641-3.

[4] H. Ma *et al.*, “Can Clinical Symptoms and Laboratory Results Predict CT Abnormality? Initial Findings Using Novel Machine Learning Techniques in Children With COVID-19 Infections.,” *Front Med (Lausanne)*, vol. 8, p. 699984, 2021, doi: 10.3389/fmed.2021.699984.

[5] J. F. Ludvigsson, “Case report and systematic review suggest that children may experience similar long-term effects to adults after clinical COVID-19.,” *Acta Paediatr*, vol. 110, no. 3, pp. 914–921, Mar. 2021, doi: 10.1111/apa.15673.

[6] A. A. Asadi-Pooya *et al.*, “Long COVID in children and adolescents.,” *World J Pediatr*, vol. 17, no. 5, pp. 495–499, Oct. 2021, doi: 10.1007/s12519-021-00457-6.

[7] B. L. Gottesman, J. Yu, C. Tanaka, C. A. Longhurst, and J. J. Kim, “Incidence of New-Onset Type 1 Diabetes Among US Children During the COVID-19 Global Pandemic,” *JAMA Pediatrics*, Jan. 2022, doi: 10.1001/jamapediatrics.2021.5801.

[8] K. K. Sum *et al.*, “COVID-19–Related Life Experiences, Outdoor Play, and Long-term Adiposity Changes Among Preschool- and School-Aged Children in Singapore 1 Year After Lockdown,” *JAMA Pediatrics*, Jan. 2022, doi: 10.1001/jamapediatrics.2021.5585.

[9] A.-L. Barabási, *Network science*. 2016.

[10] X. Liao, D. Zheng, and X. Cao, “Coronavirus pandemic analysis through tripartite graph clustering in online social networks,” *Big Data Mining and Analytics*, vol. 4, no. 4, pp. 242–251, Dec. 2021, doi: 10.26599/BDMA.2021.9020010.

[11] X. Lei and C. Zhang, “Predicting Metabolite-Disease Associations Based on Linear Neighborhood Similarity with Improved Bipartite Network Projection Algorithm,” *Complexity*, vol. 2020, p. 9342640, May 2020, doi: 10.1155/2020/9342640.

[12] S. Li, M. Xie, and X. Liu, “A Novel Approach Based on Bipartite Network Recommendation and KATZ Model to Predict Potential Micro-Disease Associations,” *Frontiers in Genetics*, vol. 10, 2019, doi: 10.3389/fgene.2019.01147.

[13] C. Fan, X. Lei, and F.-X. Wu, “Prediction of CircRNA-Disease Associations Using KATZ Model Based on Heterogeneous Networks.,” *Int J Biol Sci*, vol. 14, no. 14, pp. 1950–1959, 2018, doi: 10.7150/ijbs.28260.

[14] Q. Zhao, Y. Yang, G. Ren, E. Ge, and C. Fan, “Integrating Bipartite Network Projection and KATZ Measure to Identify Novel CircRNA-Disease Associations.,” *IEEE Trans Nanobioscience*, vol. 18, no. 4, pp. 578–584, Oct. 2019, doi: 10.1109/TNB.2019.2922214.

[15] V. J. M. Watzlaf, J. H. Garvin, S. Moeini, and P. Anania-Firouzan, “The Effectiveness of ICD-10-CM in Capturing Public Health Diseases,” *Perspect Health Inf Manag*, vol. 4, p. 6, Jun. 2007.

[16] D. Koller, G. Schön, I. Schäfer, G. Glaeske, H. van den Bussche, and H. Hansen, "Multimorbidity and long-term care dependency - A five-year follow-up," *BMC geriatrics*, vol. 14, p. 70, May 2014, doi: 10.1186/1471-2318-14-70.

[17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[18] J. M. Pujol, V. Erramilli, and P. Rodriguez, "Divide and Conquer: Partitioning Online Social Networks," *arXiv:0905.4918 [cs]*, May 2009, Accessed: Apr. 02, 2022. [Online]. Available: http://arxiv.org/abs/0905.4918

[19] J. Haynes and I. Perisic, "Mapping search relevance to social networks," 2009. doi: 10.1145/1731011.1731013.

[20] G. Roma and P. Herrera, "Community Structure in Audio Clip Sharing," in *2010 International Conference on Intelligent Networking and Collaborative Systems*, Nov. 2010, pp. 200–205. doi: 10.1109/INCOS.2010.87.

[21] D. Greene, D. Doyle, and P. Cunningham, "Tracking the Evolution of Communities in Dynamic Social Networks," Aug. 2010, vol. 2010, pp. 176–183. doi: 10.1109/ASONAM.2010.17.

[22] D. Meunier, R. Lambiotte, A. Fornito, K. D. Ersche, and E. T. Bullmore, "Hierarchical Modularity in Human Brain Functional Networks," *Front Neuroinformatics*, vol. 3, p. 37, Oct. 2009, doi: 10.3389/neuro.11.037.2009.

[23] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Sci Rep*, vol. 9, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/s41598-019-41695-z.

[24] S. H. H. Anuar *et al.*, “Comparison between Louvain and Leiden Algorithm for Network Structure: A Review,” *J. Phys.: Conf. Ser.*, vol. 2129, no. 1, p. 012028, Dec. 2021, doi: 10.1088/1742-6596/2129/1/012028.

[25] “Lipid Metabolism Disorders.” https://medlineplus.gov/lipidmetabolismdisorders.html (accessed Apr. 05, 2022).

[26] K. Adeli, J. Taher, S. Farr, C. Xiao, and G. F. Lewis, “Chapter 19 - Diabetic Dyslipidaemia,” in *Biochemistry of Lipids, Lipoproteins and Membranes (Sixth Edition)*, N. D. Ridgway and R. S. McLeod, Eds. Boston: Elsevier, 2016, pp. 549–573. doi: 10.1016/B978-0-444-63438-2.00019-5.

[27] R. Jalota Sahota and M. P. Soos, “Subclavian Vein Thrombosis,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2022. Accessed: Apr. 05, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK559269/

[28] “Mysterious wave of COVID toes still has scientists stumped,” *Science*, Mar. 31, 2022. https://www.nationalgeographic.com/science/article/mysterious-wave-of-covid-toes-still-has-scientists-stumped (accessed Apr. 05, 2022).

[29] “Cranial Nerve Disorders - Penn Medicine.” https://www.pennmedicine.org/for-patients-and-visitors/find-a-program-or-service/neurosurgery/cranial-nerve-disorders (accessed Apr. 05, 2022).