Chaos
An Interdisciplinary Journal
of Nonlinear Science

AIP Publishing

Deciphering the Molecular Mechanism of Post-Acute Sequelae of COVID-19 through Comorbidity Network Analysis

Lue Tian, MSc[1], Eric Wan, PhD[2,3,11], Sze Ling Celine Chui, PhD[4,5,11], Shirely Li, PhD[2,6,11], Esther Chan, PhD[2,11], Hao Luo, PhD[7,8], Ian C.K. Wong, PhD[2,9,11*], Qingpeng Zhang, PhD[2,10,11*]

1. School of Data Science, City University of Hong Kong, Hong Kong, China.
2. Department of Pharmacology and Pharmacy, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.
3. Department of Family Medicine and Primary Care, School of Clinical Medicine, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.
4. School of Nursing, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.
5. School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.
6. Department of Medicine, School of Clinical Medicine, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.
7. Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong, China.
8. School of Public Health Sciences, The University of Waterloo, Waterloo ON N2L3G1, Canada.
9. School of Pharmacy, Aston University, Birmingham B4 7ET, UK.
10. Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong, China.
11. Laboratory of Data Discovery for Health, Hong Kong, China.

*Correspondence: wongick@hku.hk (ICK Wong), qpzhang@hku.hk (Q Zhang)

Chaos
An Interdisciplinary Journal
of Nonlinear Science

ABSTRACT

The post-acute sequelae of COVID-19 (PASC) poses a significant health challenge in the post-pandemic world. However, the underlying biological mechanisms of PASC remain intricate and elusive. Network-based methods can leverage electronic health record (EHR) data and biological knowledge to investigate the impact of COVID-19 on PASC and uncover the underlying biological mechanisms. This study analyzed territory-wide longitudinal electronic health records (from January 1, 2020, to August 31, 2022) of 50296 COVID-19 patients and a healthy non-exposed group of 100592 individuals to determine the impact of COVID-19 on disease progression, provide molecular insights, and identify associated biomarkers. We constructed a comorbidity network and performed disease-protein mapping and protein-protein interaction network analysis to reveal the impact of COVID-19 on disease trajectories. Results showed disparities in prevalent disease comorbidity patterns, with certain patterns exhibiting a more pronounced influence by COVID-19. Overlapping proteins elucidate the biological mechanisms of COVID-19's impact on each comorbidity pattern, and essential proteins can be identified based on their weights. Our findings can help clarify the biological mechanisms of COVID-19, discover intervention methods, and decode the molecular basis of comorbidity associations, while also yielding potential biomarkers and corresponding treatments for specific disease progression patterns.

Keywords: network science, comorbidity, network medicine, post-acute sequelae of COVID-19

Word count: 4513 (5598 including figures and tables)

The accurate identification of comorbidity patterns associated with elevated COVID-19 infection risk is essential for effective medical resource allocation, prioritizing care, and supporting patient recovery. This study delved into the biological mechanisms underlying post-acute sequelae of COVID-19 by analyzing the comorbidity network. Key proteins and significant biological pathways were identified through Protein-Protein Interaction network analysis and Gene Ontology enrichment analysis. These insights not only contribute to understanding the fundamental mechanisms of post-acute sequelae but also hold potential as biomarkers and therapeutic targets, laying the groundwork for the development and repurposing of drugs to benefit COVID-19 patients.

INTRODUCTION

The global pandemic of Coronavirus Disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, has impacted billions of people, resulting in millions of deaths, and leaving tens of millions suffering from persistent symptoms and signs after the acute phase of COVID-19[1]. This phenomenon, referred to as post-acute sequelae of COVID-19 (PASC), is gradually coming to an end[2,3]. PASC exhibits heterogeneous manifestations and severity[4], impacting various organ systems including cardiovascular[5,6], mental[7], metabolic[8], and renal systems[9]. However, the underlying biological mechanisms of PASC remain intricate and elusive.

Current research on the potential biological mechanisms of COVID-19 and PASC primarily involves small patient cohorts and concentrates on the relationship between COVID-19 and diseases in specific systems or organs individually[2,10–12]. Investigations involving large patient cohorts and associations between COVID-19 and diseases across multiple organs and systems can aid in uncovering the biological mechanisms of multimorbidity present in PASC and pre-existing diseases influenced by COVID-19. Comorbidity and multimorbidity[13] refer to the co-occurrence of two or more diseases in an individual. If the frequency of co-occurrence of diseases exceeds the frequency of disease combinations selected by chance, multimorbidity exists among these diseases. Research on multimorbidity associations with PASC has shown that pre-existing multimorbidity may drive PASC[10,14]. The underlying mechanisms of multimorbidity are complex, potentially involving shared genetic or

Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing

environmental factors or resulting from the treatment or intervention for one disease leading to the development of another[14]. Studies on multimorbidity relations in PASC have considered the influence of demographic factors[15,16] like sex, age, race, COVID-19 vaccine injection, and patient electronic health records (EHR) diseases. However, the underlying biological mechanisms of these multimorbidity relations influenced by COVID-19 remain unclear.

Our research utilized territory-wide EHR data from the Hong Kong Hospital Authority to investigate the impact of COVID-19 on PASC. We employed network-based methods to assess the influence of COVID-19 on multimorbidity across multiple organs and systems in PASC. Individuals with specific pre-existing diseases may have a higher risk of developing certain diseases included in PASC due to the effects of COVID-19. To explore the underlying mechanisms of these particular multimorbidity relations, we incorporated biological knowledge from the protein-protein interaction (PPI) network[17] and Gene Ontology (GO) enrichment analysis[18] to identify the most affected and essential proteins that could serve as potential targets for future interventions, such as preventive measures.

METHODS

Study design and population

All electronic datasets included in this research are from the Hong Kong Hospital Authority (HKHA) database. Based on the COVID-19 record (based on rapid antigen test [RAT] or polymerase chain reaction [PCR] test in throat swab, nasopharyngeal aspirate, or deep throat sputum specimens), patients are divided into two groups: exposure and non-exposed groups. (Supplementary Figure 4)

For the exposure group, the diagnose records within 730 days before COVID-19 and 180 days after COVID-19 are retained. We then exclude all diagnose records within 28 days (acute-phase of infection) after COVID-19. For each individual, diseases appearing before COVID-19 were considered as pre-existing diseases, and new emerging diseases appearing after COVID-19 were considered as post-infection diseases.

For the non-exposed group, we applied the same procedure to identify the pre-existing and post-infection diseases for each individual by treating the date 180 days before the last record date as the simulated COVID-19 infection date.

To investigate the influence of COVID-19 on PASC, we compared the differences in comorbidity patterns between the exposure group (people with first COVID-19 in 2022) and the non-exposed group (people without COVID-19). To ensure a fair comparison, we employed propensity score matching[19] to select a non-exposed group with pre-existing disease records similar to the exposure group. In the matching process, we considered not only individual EHR data denoted by 3-digital ICD-9 codes from the baseline period, but also demographic data (age, sex) and vaccine information (vaccine number). After matching, we calculated the standardized mean difference (SMD) to quantify the balance for each confounder. An SMD value below 0.1 serves as a threshold to determine whether the confounder is well-balanced.

Propensity Score Matching

We included all disease related 3-digital ICD-9 CM codes from pre-existing diseases in both non-exposed and exposure groups. Each appearing ICD-9 code serves as a feature in the Propensity Score Matching. The number of vaccinations received prior to COVID-19 is included as a feature. For each individual in the exposure group, we select the two nearest neighbors from the non-exposed group based on the processed propensity score, which is obtained by applying a logit function. After propensity score matching, the COVID-19 and non-exposed groups include 50296 and 100592 patients respectively, from 58753 in the COVID-19 group and 488670 in the non-exposed group. (details are shown in Supplementary Table 2) In the following comorbidity patterns' coefficient computation part, we considered patients with at least one new disease and excluded patients without new post-infection diseases.

Comorbidity network construction

According to COVID-19 infection date for each individual in the exposure group (simulated for the non-exposed group). We first define pre-existing diseases as those diagnosed before COVID-19, and

post-infection diseases as new emerging diseases diagnosed >28 days after COVID-19. We generated disease pairs by selecting the first disease from all pre-existing diseases and the second disease from all post-infection diseases and applied Pearson correlation coefficient and relative risk to quantify the co-occurrence of two diseases composing each disease pair. [20] To reduce bias from data and computation methods, we used both values to identify co-occurring disease pairs combining with chi-square and fisher exact test. [20] We identified disease pairs appearing in the exposure group with Pearson correlation coefficient > 0 and relative risk > 1 as comorbidity patterns. To investigate comorbidity patterns influenced by COVID-19, we selected comorbidity patterns according to the three requirements: (1) comorbidity patterns are significantly more frequent (Fisher exact test or chi-square test, p-value < 0.05) in the exposure group than in the non-exposed group. (2) the count number of each comorbidity pattern in the exposure group is at least 10. (3) the Pearson correlation coefficient and relative risk of each comorbidity are larger in the exposure group than in the non-exposed group. Selected comorbidity patterns are used for comorbidity network construction. In the comorbidity network, each node represents a disease, and each directed edge from a pre-existing disease to a post-infection disease represents a comorbidity pattern indicating higher prevalence in the exposed group than in the non-exposed group. For instance, an edge like (disease A, disease B) suggests that individuals with pre-existing disease A face an elevated risk of developing disease B following COVID-19, compared to those without COVID-19 exposure. The difference in Pearson correlation coefficients and relative risk for each comorbidity pattern between exposure group and non-exposed group are considered as additional attributes for edges.

Relative Risk denoted by RR_{ij} represents the relative risk between disease i and disease j. C_{ij} is the number of co-occurrence incidences of disease i and disease j, I_i is the number of incidences of disease i, I_j is the number of incidences of disease j, and N is the number of patients included in the dataset (for the non-exposed group and exposure group).

$$RR_{ij} = C_{ij}/\left(I_i I_j/N\right) \quad (1)$$

Pearson correlation is another common method to evaluate the strength of diseases' connection. The Pearson correlation between disease i and disease j is denoted by ϕ_{ij}, and the formula is the following:

$$\phi_{ij} = \left[\left(N\ C_{ij}\right) - \ I_i I_j\right]/sqrt\left(I_i I_j (N - \ I_i)\left(N - \ I_j\right)\right) \quad (2)$$

Disease pairs with Pearson correlation coefficient > 0 and relative risk > 1 imply that these diseases are more likely to occur than by chance.

Protein-Protein Interaction (PPI) network and SARS-CoV-2 human proteins

The protein-protein interaction network used in this study was assembled from 21 public databases by Barabási[21]. The final interactome used in our study contains 18,505 proteins and 327,924 interactions between them. For SARS-CoV-2 human proteins, we used related data detected by Gordon[22]. To quantify the distance between the post-infection and pre-existing diseases, we utilized node-node distances for each protein pair on the largest connected component of the protein-protein interaction network, which contains 18446 nodes and 327868 edges. All proteins are represented by their encoded genes (Entrez ID and Symbol ID).

ICD code and Gene/Protein association data

Data were derived from the DisGeNET[23,24] database and the OMIM[25,26] dataset. These datasets encompass information about proteins and diseases, as well as their interrelationships. Utilizing these datasets, we were able to identify proteins associated with each disease. The distances between diseases were determined based on the distances between their associated proteins. From the DisGeNET database and the OMIM dataset, we assigned associated proteins to 537 diseases, each disease identified by an ICD-9 code and each disease-associated protein identified by its encoded gene (Entrez ID and Symbol ID).

Distance Measure

Disease-protein relations are utilized to map each disease, denoted by an ICD-9 code, to the protein-protein interaction (PPI) networks. Diseases are represented by protein groups, which are sets of proteins associated with a specific disease. The topological distance between corresponding protein

groups in the PPI networks measures the distance between two diseases. A shorter distance between the protein groups indicates a closer relation between the diseases.

For example, if we have a pre-existing disease A and a post-infection disease B, we can map them to the PPI networks and find their corresponding protein groups. Then, we can calculate the distance between them as d_{AB}, where $d(a,b)$ represents the shortest path from protein a to protein b in the PPI networks.

$$\langle d_{AB} \rangle = \frac{1}{\|B\|} \sum_{b \in B} \min_{a \in A} d(a,b) \tag{3}$$

To compute the distance between a pre-existing disease A with COVID-19 and a post-infection disease B, we need to integrate proteins associated with COVID-19 (denoted by C) and proteins associated with disease A. The computation formula is as follows:

$$\langle d_{AB}^{C} \rangle = \frac{1}{\|B\|} \sum_{b \in B} \min_{a \in A \cup C} d(a,b) \tag{4}$$

The distance change (denoted by Δd_{AB}) for a pre-existing disease A and a post-infection disease B with and without the addition of proteins associated with COVID-19 is as follows:

$$\Delta d_{AB} = \langle d_{AB} \rangle - \langle d_{AB}^{C} \rangle \tag{5}$$

We also apply the following methods to compute disease distances for sensitivity analysis:

$$\langle d_{AB} \rangle = \frac{1}{\|A\|\|B\|} \sum_{b \in B} \sum_{a \in A} d(a,b) \tag{6}$$

Testing Method

We employed two statistical tests, the Chi-square test and the Fisher's exact test, to compare the co-occurrence frequency of disease pairs between the non-exposed and exposure groups. These tests can assist us in determining whether a significant association exists between two diseases in the presence or absence of COVID-19.

Among all 537 diseases which have at least one associated protein, for each pre-existing disease, we treated selected comorbidity patterns as positive samples and generated negative samples (disease pairs composed of pre-existing disease and other diseases, which are different from selected comorbidity patterns). Additionally, we utilized the Wilcoxon signed-rank test to compare the changes of protein-

protein interaction (PPI) distance before and after the addition of proteins associated with COVID-19 to proteins associated with pre-existing disease between each disease pair in the positive group and all disease pairs with the same pre-existing disease in the negative group. This test can help us evaluate whether a significant difference exists in the PPI distance between each disease pair in the positive group and all disease pairs with the same pre-existing disease in the negative sample group, thereby indicating the impact of COVID-19 on the relationship between two diseases.

Negative Sample for Z-score Compute

For each pre-existing post-infection disease pair, we conduct a permutation test of 1000 repeats to compute Z-score. We randomly select proteins from the PPI network, and the selected protein groups are required to have similar node degree distribution to pre-existing disease and post-infection disease respectively. Distance differences before and after adding proteins associated with COVID-19 as a part of pre-existing disease associated proteins are also calculated based on the randomly selected protein groups, and the mean value and standardized deviation of the result are used to compute Z-score for each pre-existing post-infection disease pair according to the following formula:

$$Z_{\Delta d_{AB}} = \frac{\Delta d_{AB} - \mu_r}{\sigma_r} \tag{7}$$

Δd_{AB} is the distance change of pre-existing post-infection disease pair before and after adding proteins associated with COVID-19 as a part of pre-existing disease associated proteins. μ_r is the mean value of the distance change from the permutation test, σ_r is the standardized deviation of the distance change from the permutation test.

GO enrichment Analysis

GO terms describe the functions of gene products across three primary aspects: biological process, molecular function, and cellular component. By conducting GO enrichment analysis, we can pinpoint the GO terms and crucial genes most impacted by COVID-19, thereby enhancing our understanding of the disease's underlying biology.[18,27–29]. We employed Fisher's exact test for the enrichment analysis and used the Benjamini-Hochberg procedure to adjust the p-values for multiple testing.[30,31].

GO Term and Gene Information are from the National Center for Biotechnology Information(NCBI). These datasets include information on GO terms and the relationships between GO terms, genes, and proteins. We can also identify proteins associated with each GO term. By utilizing GO enrichment analysis, we can discover highly influenced GO terms for each disease based on its associated proteins.

Protein and GO terms Evaluation

The importance of overlapping proteins and GO terms (those associated with both the post-infection disease and pre-existing disease with COVID-19) is the sum of the coefficients of disease pairs they are associated with based on TF-IDF metric. [32] We can sort overlapping proteins and GO terms by Relative Risk (RR), Correlation Coefficient and frequency, respectively. The final index of each item is determined by the mean index of those three indices. According to the final index, we can identify important proteins and GO terms.

RESULTS

Overall pipeline

Using propensity score matching, we construct a non-exposed group (patients without COVID-19 record) and an exposure group (COVID-19 patients) with similar clinical records prior to their respective first COVID-19 record (simulating infection date for healthy individuals) from January 1, 2020 to August 31, 2022. Then, we construct and analyze the comorbidity network as follows:

- For each individual, we first define *pre-existing diseases* as those diagnosed within 730 days before COVID-19, and *post-infection diseases* as new diseases diagnosed in 28 days to 180 days after COVID-19. We define the *disease pair* as a pair of a pre-existing disease and a post-infection disease in the exposure group. *Comorbidity patterns* refer to these disease pairs with a positive Pearson correlation coefficient between pre-existing and post-infection diseases with a relative risk > 1 (More details are shown in the Method part).
- We then construct a comorbidity network[33] consisting of these comorbidity patterns and involved diseases. In our comorbidity network, each node represents a disease. The existence of a directed edge from the pre-existing disease to the post-infection disease indicates the comorbidity pattern is significantly more frequent (p-value < 0.05, larger Pearson correlation coefficient and relative

Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing

risk) in the exposure group than in the non-exposed group. The difference in Pearson correlation coefficient and relative risk for each comorbidity pattern between the exposure group and the non-exposed group are considered as additional attributes for edges.

- For each comorbidity pattern, as represented by an edge in the comorbidity network, network analysis and GO enrichment[18,34,35] analysis were employed for biological pathways discovery by utilizing corresponding disease-associated proteins with COVID-19 associated proteins, and important proteins and GO terms were identified.

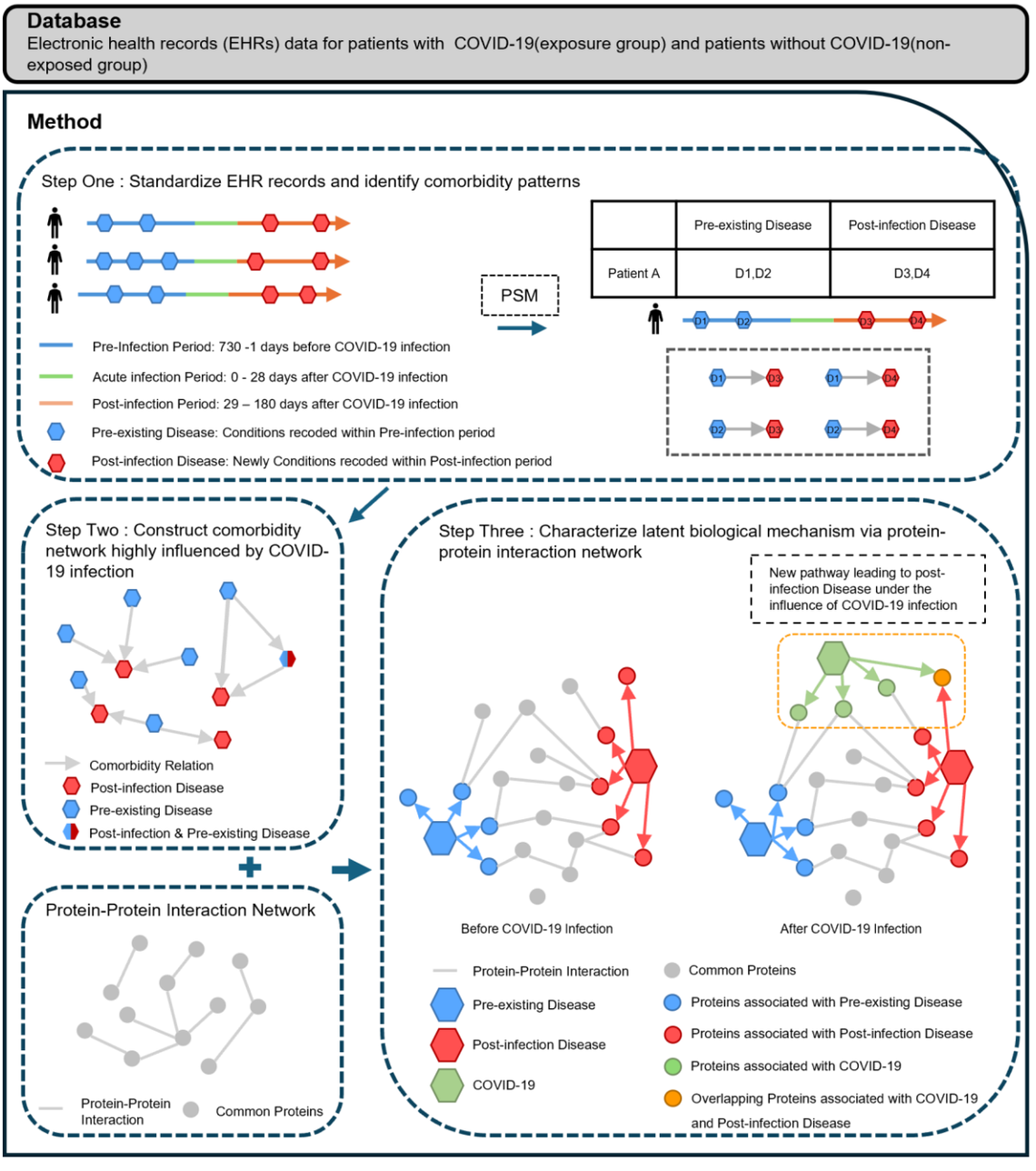

Figure 1. Data curation and analysis pipeline. To explore the molecular mechanism of Post-Acute Sequelae of COVID-19, we stratify patients to the exposure group and the non-exposed group and define disease pairs based on their electric health records after propensity score matching (step 1). After filtering disease pairs, we construct a comorbidity network based on the correlation coefficient between pre-existing and post-infection diseases (step 2). Finally, we utilize the protein-protein interaction network to identify potential key proteins and biological pathways in comorbidity patterns under the influence of COVID-19 (step 3).

Study cohorts

Before matching, the dataset comprised 58753 individuals with COVID-19 and 488670 individuals without COVID-19 (refer to Supplementary Table 1). Following the matching process (as detailed in the Methods section), the exposure group contained 50296 observations, while the non-exposed group contained 100592 observations. The median age for the non-exposed group is 65, compared to 66 for the exposure group. The non-exposed group was composed of 52.4% males and 47.6% females, whereas the exposure group consisted of 52.2% males and 47.8% females. Over 50% of the individuals in both groups are aged 60 years or older. Further details regarding specific diseases are shown in Supplementary Figure 1 and the SMD of features are shown in Supplementary Table 2. Utilizing the electronic health records of each individual in the two groups, we were able to construct a directed comorbidity network, which illustrates the disease trajectories before and after COVID-19.

Table 1. Summary statistics of the dataset.

		Overall	Non-exposed group	exposure group
Num		150888	100592	50296
Vaccine num, n (%)	**0 dose**	46013 (30.5)	30880 (30.7)	15133 (30.1)
	1 dose	22015 (14.6)	14564 (14.5)	7451 (14.8)
	2 doses	68563 (45.4)	45705 (45.4)	22858 (45.4)
	3 doses	10903 (7.2)	7248 (7.2)	3655 (7.3)
	4 doses	3394 (2.2)	2195 (2.2)	1199 (2.4)
Age, median [Q1,Q3]		65.0 [53.0,76.0]	65.0 [53.0,76.0]	65.0 [53.0,76.0]
Sex, n (%)	**Female**	78933 (52.3)	52682 (52.4)	26251 (52.2)
	Male	71955 (47.7)	47910 (47.6)	24045 (47.8)
Num of diseases before COVID-19, mean (SD)		4.0 (3.9)	4.0 (4.0)	3.9 (3.7)

Num of diseases after COVID-19, mean (SD)		1.8 (1.9)	1.9 (2.0)	1.6 (1.8)
Num of pre-existing diseases before COVID-19 for exposure group and Non-exposed group (simulate), n (%)	**1 disease**	42406 (28.1)	28405 (28.2)	14001 (27.8)
	2 diseases	30099 (19.9)	20499 (20.4)	9600 (19.1)
	3 diseases	20287 (13.4)	13349 (13.3)	6938 (13.8)
	4 diseases	13823 (9.2)	8951 (8.9)	4872 (9.7)
	more than 5 diseases	44273 (29.3)	29388 (29.2)	14885 (29.6)
Num of post-infection diseases after COVID-19 for exposure group and Non-exposed group (simulate), n (%)	**1 disease**	64244 (42.6)	42312 (42.1)	21932 (43.6)
	2 diseases	27835 (18.4)	18907 (18.8)	8928 (17.8)
	3 diseases	12900 (8.5)	8898 (8.8)	4002 (8.0)
	4 diseases	7169 (4.8)	4956 (4.9)	2213 (4.4)
	more than 5 diseases	12208 (8.1)	8863 (8.8)	3345 (6.7)
	no disease	26532 (17.6)	16656 (16.6)	9876 (19.6)
Age group, n (%)	**0-20**	4228 (2.8)	2927 (2.9)	1301 (2.6)
	20-40	14389 (9.5)	9691 (9.6)	4698 (9.3)
	40-60	35999 (23.9)	23778 (23.6)	12221 (24.3)
	60-80	65776 (43.6)	43552 (43.3)	22224 (44.2)
	80+	30496 (20.2)	20644 (20.5)	9852 (19.6)

Comorbidity Network Analysis

Our comorbidity network comprises 96 nodes and 161 edges, representing the comorbidity patterns significantly influenced by COVID-19. Nodes are categorized into 14 disease groups according to their ICD9 categories. Edges are classified into two groups according to the adjacent nodes: intra-group edges (source node and target node in the same disease group), inter-group edges (source node and target node in different disease groups).

Figure 2 (A) (B) depict the constructed comorbidity network and disease group classifications. The network is heterogeneously connected, with most disease groups sparsely associated with other disease groups and a few disease groups more closely related to some other disease groups. More specifically, among all groups, disease group (001–139 Infectious and Parasitic Diseases) and disease group (460–519 Diseases of the Respiratory System) have a higher frequency than other disease groups. 079 (Viral and chlamydial infection in diseases classified elsewhere and of unspecified site) and 519 (Other diseases of respiratory system) are two most frequent diseases among all, indicating that COVID-19 has the most significant impact on the respiratory system. Additionally, the neural, gastrointestinal and circulatory systems are also frequently affected by the COVID-19, which is aligns with the literature.

[36–38] Diseases such as 276 (Disorders of fluid electrolyte and acid-base balance), 428 (Heart failure), 788 (Symptoms involving urinary system), 272 (Disorders of lipoid metabolism), 294 (Persistent mental disorders due to conditions classified elsewhere) are also more likely to be involved in the comorbidity relationships. (Refer to Supplementary Table 3)

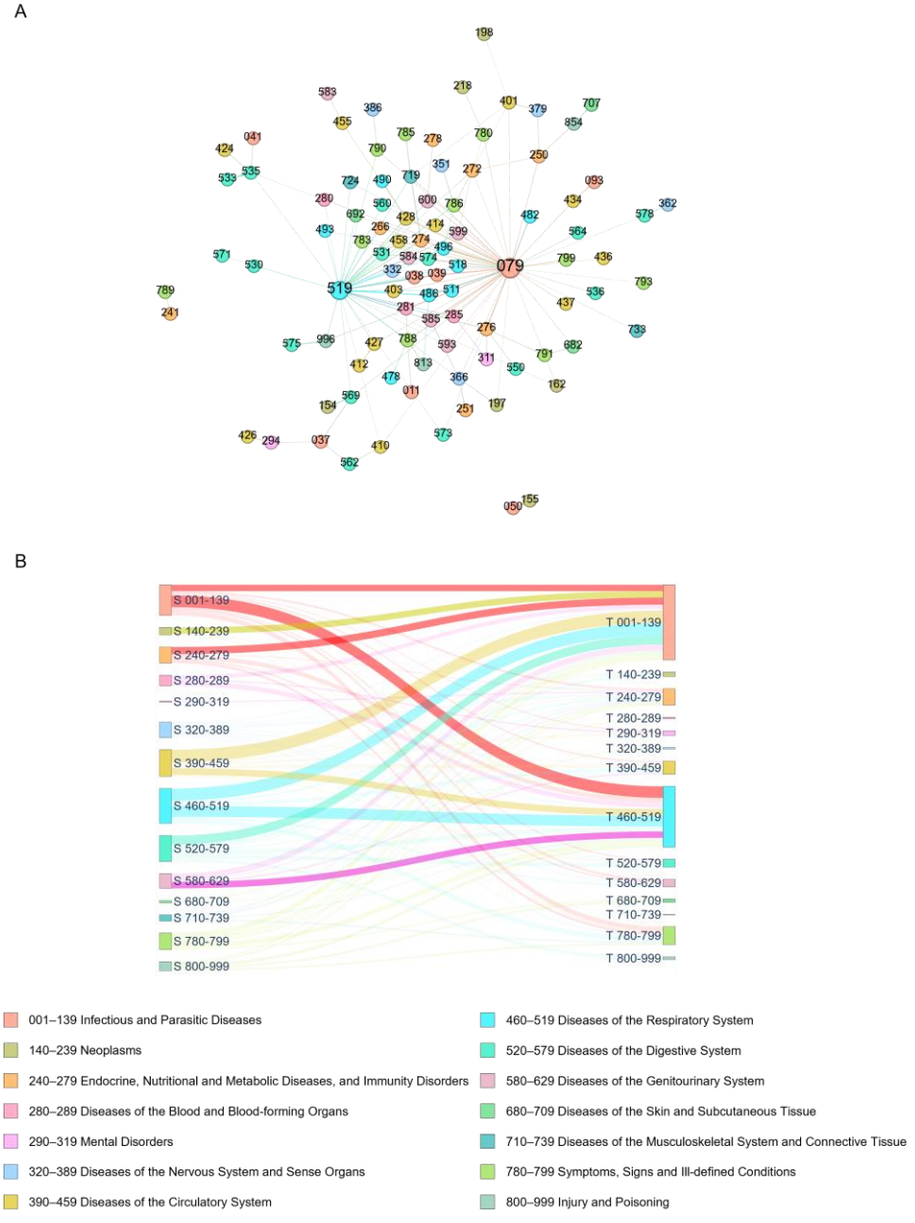

Figure 2. Visualization of comorbidity patterns. (A) The comorbidity network. (B) The associations among disease groups. In (A), the size of a node is proportional to its occurrence in our dataset. In (B), the rectangles in the Sankey diagram correspond to ICD 9 disease categories. The left rectangles

represent disease groups of pre-existing diseases. The right rectangles represent disease groups of post-infection diseases. The edge linking a pair of rectangles indicates that the comorbidity patterns are significantly more frequent among COVID-19 patients as compared to those in the non-exposed group. The thickness of an edge is proportional to increased occurrence among COVID-19 patients as compared to those in the non-exposed group. Please refer to Supplementary Table 5 for more details.

The comorbidity network's edges suggest more pronounced comorbidity relationships in the exposure group compared to the non-exposed group. These relationships imply that patients with a history of respiratory system diseases are at an increased risk of developing further respiratory system diseases due to COVID-19. Additionally, these patients with a history of respiratory system diseases also demonstrate a heightened risk for diseases within the circulatory system, genitourinary system, among others. Patients previously diagnosed with essential hypertension also exhibit a higher risk of developing respiratory system diseases and lipid disorders due to COVID-19. Patients with a history of peptic ulcer (site unspecified) are at an increased risk of developing Gastritis and duodenitis (Refer to Supplementary Table 4).

Biological Mechanism Explanation

To investigate the biological mechanisms underlying identified comorbidity relations, we utilize the PPI and GO terms associated with the diseases. We hypothesize that for COVID-19 to influence the disease comorbidity patterns of patients, its host factors (genes/proteins) should be localized in the corresponding subnetwork within the human PPI network, either directly targeting the disease-associated genes/proteins or indirectly affecting them through PPIs. Specifically, we hypothesize that these comorbidity patterns identified from the comorbidity network experience a larger reduction in the topological distance in the PPI network caused by the inclusion of proteins associated with COVID-19 as additional proteins associated with pre-existing disease. Similar patterns were previously observed in the relationship between COVID-19 and brain microvascular injury. [39]

After we eliminated disease pairs without associated protein information, 145 pre-existing disease-post-infection disease pairs, including 84 disease types (74 pre-existing diseases, 37 post-infection diseases), remained. Then, the network distance from pre-existing diseases to post-infection diseases is measured on the PPI network. The network distance is measured as the average shortest path length between each protein associated with post-infection disease and proteins associated with pre-existing disease. (Equation 3) Next, we measure the change of network distance of pre-existing diseases and post-infection diseases in the PPI network when treating COVID-19 as an additional pre-existing disease (Equation 4, Equation 5).We observed that the network distance between the disease pairs with elevated comorbidity risk after COVID-19 became significantly shorter because of the addition of proteins associated with COVID-19 in 54 disease pairs among 145 disease pairs (Figure 3). Please refer to the Method section for details. The result of Wilcoxon signed-rank test about distance change between exposure and non-exposed groups is shown in Figure 3 (A), Supplementary Figure 2 and Supplementary Table 6. The result of Z scores about the distance change within exposure groups is shown in Figure 3 (B), Supplementary Figure 3 and Supplementary Table 7. The sensitivity analysis is shown in Supplementary Figure 5,6,7 and Supplementary Table 12,13,14,15.

Further GO analysis of involved proteins reveals that COVID-19 introduced additional mechanistic pathways towards post-infection diseases, effectively increasing the risk of developing post-infection diseases. Please refer to Supplementary Table 8, 9 for a list of frequent GO terms associated with these proteins.

Among the involved proteins, *overlapping* proteins (those associated with both the post-infection disease and pre-existing disease with COVID-19) play a major role in shortening the distance between the disease pairs. The distance of each disease pair reflects the likelihood that proteins associated with post-infection disease are influenced by the abnormal expression of proteins associated with pre-existing disease with COVID-19 via the PPI network. These proteins representing biological functions are linked to the phenotype of post-infection disease. For example, in Figure 4, the overlapping proteins associated with both the post-infection disease (272) and COVID-19 all involved in lipid metabolism,

the abnormal expression of the proteins may lead to a perturbance of lipid metabolism related biological functions resulting in the phenotype Disorders of lipoid metabolism[40]. Please refer to the Supplementary Materials for a list of the most frequent overlapping proteins (Refer to Table 2, Supplementary Table 10, 11). GO enrichment analysis reveals that, in addition to the roles in COVID-19 and following inflammatory response, the expression disorder of these overlapping proteins is leading towards other diseases involving cardiovascular system, urinary system, and respiratory system.

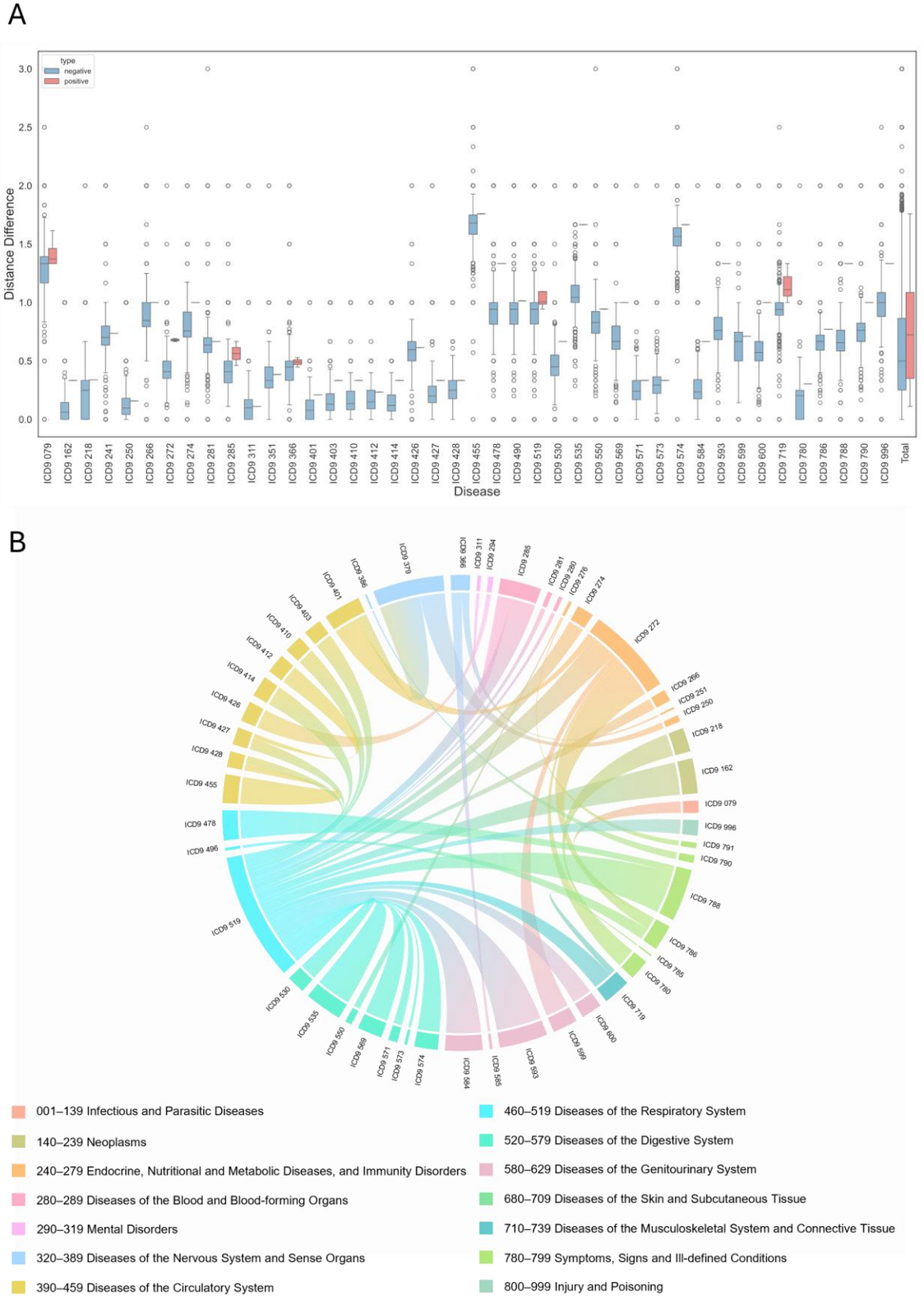

Figure 3. Visualization of PPI distance. (A) The distribution of the distance changes from the corresponding disease to other comorbidities in the positive sample and negative sample (including

significant ($p<0.05$) disease pairs in Wilcoxon signed-rank test). (B) Z score of comorbidity pairs (including significant disease pairs in T test). The bar is thicker, and the Z score is larger. In (A), red (blue) rectangles represent pre-existing disease groups in the positive (negative) sample. In (B), each bar in the plot represents a comorbidity pair with the wider segment part mapping to the pre-existing disease in a disease pair, while the narrower part mapping to the post-infection disease.

Table 2. Top 10 most important overlapping proteins.

Symbol ID	Type	Associated biological functions
ACE	Enzyme	Involving in blood pressure regulation and electrolyte balance
PDYN	Preproprotein	Proteolytically processing to form the secreted opioid peptides beta-neoendorphin, dynorphin, leu-enkephalin, rimorphin, and leumorphin.
GNAS	Protein	Playing a key role in the classical signal transduction pathway linking receptor-ligand interactions with the activation of adenylyl cyclase and a variety of cellular responses
GSTP1	Enzyme	Playing an important role in detoxification by catalyzing the conjugation of many hydrophobic and electrophilic compounds with reduced glutathione
IL1B	Cytokine	Acting as an important mediator of the inflammatory response, and involving in a variety of cellular activities, including cell proliferation, differentiation, and apoptosis
SCGB1A1	Secreted Proteins	Involved in numerous functions including anti-inflammation, inhibition of phospholipase A2 and the sequestering of hydrophobic ligands
PLAT	Protease	Converting the proenzyme plasminogen to plasmin, a fibrinolytic enzyme
NOS3	Enzyme	Playing an important role in detoxification by catalyzing the conjugation of many hydrophobic and electrophilic compounds with reduced glutathione
HMOX1	Enzyme	Involved in heme catabolism, cleaving heme to form biliverdin
GDF15	Ligand	Acting as a pleiotropic cytokine and participateing in the stress response program of cells after cellular injury. Increased protein levels are associated with disease states such as tissue hypoxia, inflammation, acute injury and oxidative stress.

Figure 4 illustrates the interplays between a representative disease comorbidity pair, from 401 (Essential hypertension) to 272 (Disorders of lipoid metabolism), which has an elevated risk because of the COVID-19 infection. We find that the overlapping proteins (orange and purple circles) associated with

both the post-infection disease (272) and pre-existing diseases (401 and COVID-19) play an important role in the development of disease comorbidity. There are five additional overlapping proteins (orange circles) because of the COVID-19 infection: NEU1, INHBE, NPC2, AGPS and GLA.[41,42] Specifically, NEU1 has a significant effect on lipid metabolism and inflammatory processes and is a potential drug target for decreasing atherosclerosis. [43,44] INHBE activates energy expenditure through brown/beige adipocyte activation, and it can be a potential drug target for obesity therapy. [45,46] NPC2 is essential for the pathways involved in glucose and lipid metabolism, helping the egress of lipids from the lysosome. [47,48] AGPS is an ether lipid generating enzyme which is important for the balance of structural and signaling lipids.[49] GLA is a polyunsaturated fatty acid that can reduce lipid deposition.[40] The addition of these five overlapping proteins representing biological functions related to lipid metabolism potentially explains the mechanism of how COVID-19 elevated the risk of developing the post-infection disease (272).

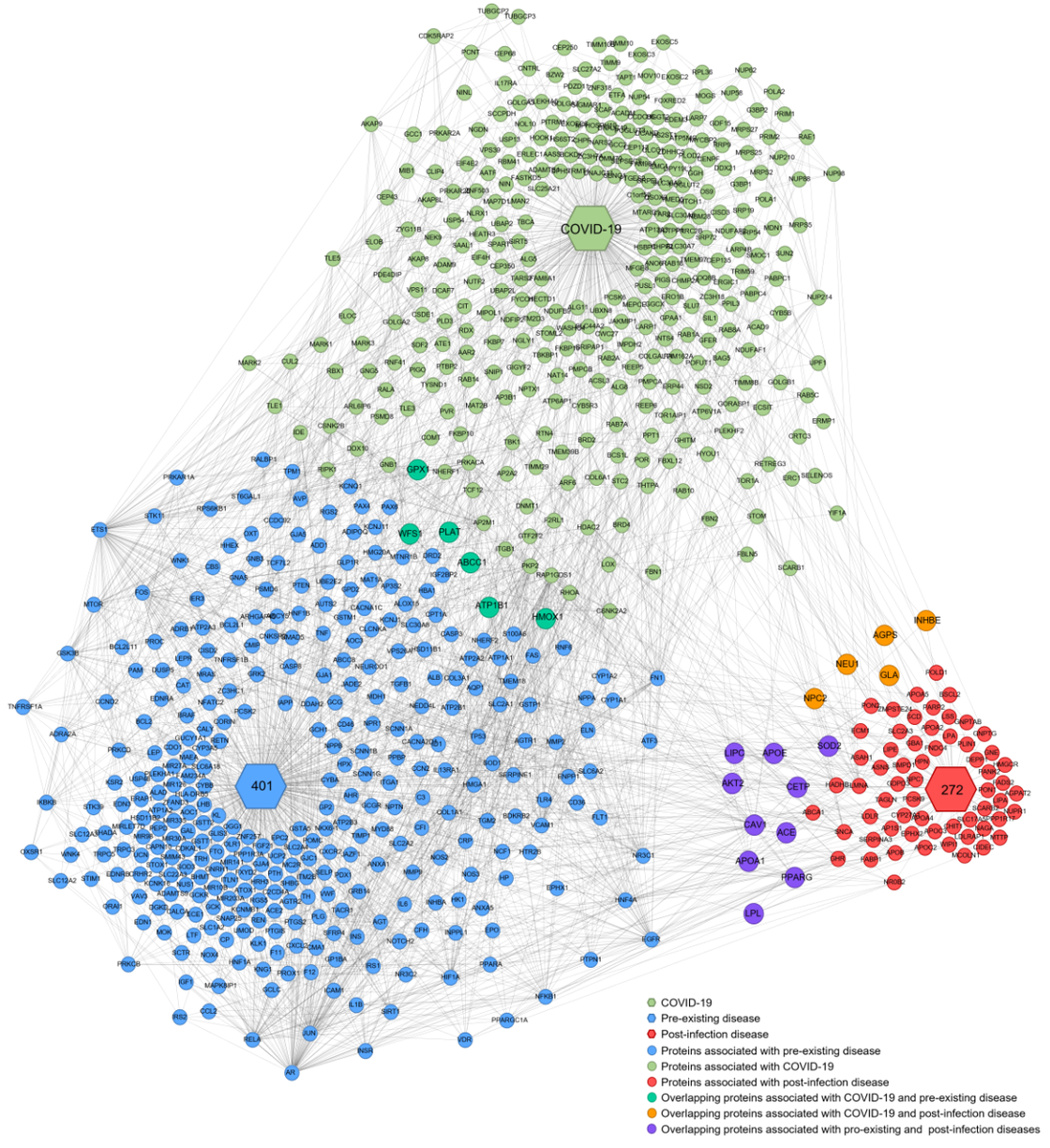

Figure 4. PPIs underlying the comorbidity pattern between ICD 401 and ICD 272. A hexagon represents a disease. A circle represents a protein. An edge between circles represents the existence of protein-protein interaction. An edge between a hexagon and a circle represents the protein associated with the disease.

Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing

DISCUSSION

Our study, leveraging population-based EHR and a wealth of biomedical data, stands as a pioneering quantitative analysis of the complex molecular mechanisms underlying the comorbidity patterns associated with COVID-19. This research is not merely an exploration but a comprehensive examination of the data, aiming to unravel the progression of the disease and the evolution of comorbidity patterns resulting from a COVID-19 infection.

Our findings provide a deeper understanding of the phenomenon known as PASC; a disease characterized by lingering symptoms after recovery from the acute phase of COVID-19 that has been a global concern for healthcare professionals. Our research illuminates the elevated-risk comorbidity patterns associated with PASC, significantly contributing to the existing body of knowledge on this subject.

Central to our study are the key proteins we identified, which play a pivotal role in increasing the risk of these comorbidity patterns. These proteins are not merely markers but potential targets for therapeutic intervention, laying the groundwork for the development of new drugs and the repurposing of existing ones. (More details are shown in Supplementary Table 10, 11) The ultimate goal is not only to reduce the risk of COVID-19 re-infection but also to prevent the onset of PASC, offering hope to millions of patients worldwide.

The practical applications of our study are extensive. Using the comorbidity patterns, we discovered and the wealth of data from electronic health records, we can identify patients who are at high risk for PASC. This information is crucial for the effective allocation of medical resources, ensuring prompt care for those who need it most. Moreover, it aids in the recovery process of patients from COVID-19, providing a roadmap for their journey back to health. Our study, therefore, stands at the intersection of research and real-world application, contributing to the fight against this global pandemic.

Our study has limitations. First, our analyses were based on the topology of the PPI network. The PPI network serves as a "skeleton" of the biological signaling circuitry in the human body. However, PPI

network cannot fully represent the pharmacokinetics and pharmacodynamics (PK/PD) associated with drugs. Future research is needed to incorporate the PK/PD models for a better understanding of the effects of drugs on the human body. Secondly, our population EHR data was obtained from public hospitals in Hong Kong. Although our data is among the most complete for a population, there is an inevitable under-reporting problem, especially for young patients. Third, although we have tried our best to build a comprehensive mapping between diseases and proteins in the PPI network. The mapping may still be subject to bias because of the lack of such data. Further biological research is needed to enrich and complement existing databases. The causative relationship between proteins in molecular mechanisms remains elusive. Validating these connections necessitates strategic planning of randomized controlled trials (RCTs) or the application of Mendelian randomization. Additionally, advanced methodologies in social network[50] and gene regulation network analyses[51–53], coupled with the acquisition of supplementary biological datasets, are essential for substantiating these causal connections.

Conclusions

In conclusion, our study significantly advances the understanding of the intricate molecular mechanisms and comorbidity patterns associated with COVID-19. By leveraging extensive population-based electronic health records and biomedical data, we have provided a comprehensive analysis that elucidates the progression of the disease and the evolution of comorbidity patterns resulting from a COVID-19 infection. Our findings identify the critical role of specific proteins in increasing the risk of these comorbidity patterns, identifying them as potential targets for therapeutic intervention. This paves the way for developing new treatments and repurposing existing drugs, ultimately aiming to reduce the risk of COVID-19 re-infection and prevent the onset of PASC. The practical implications of our research are extensive. By identifying high-risk patients through comorbidity patterns and electronic health record data, we can ensure a more effective allocation of medical resources and provide timely care to those most in need.

Supplementary materials

The supplementary materials contain additional figures and tables of the analytical results.

Ethics approval and consent to participate

Ethical approval for this study was granted by the Institutional Review Board of the University of Hong Kong/HA HK West Cluster (UW20-556, UW21-149 and UW21-138).

Consent for publication

All authors provide consent for the publication of this manuscript.

Competing interests

The authors declare no competing interests.

Availability of data and materials

Data and raw code are available at https://github.com/LueTian/LONG-COVID. Clinical data could not be deposited for patient privacy reasons.

Funding

Research Grants Council of Hong Kong.

Acknowledgements

N/A

Authors' Contributions

T.L.: methodology, data curation, formal analysis, preparation of the original draft, visualization, data interpretation; **QP**: methodology, data interpretation, writing-review and editing, conceptualization of the project, supervision, acquisition of resources; **EW, SLCC, SL**: writing-review, acquisition of

Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing

resources; **EC, HL**: acquisition of resources; **L.C.K.W**: conceptualization of the project, acquisition of resources.

Abbreviations

PASC: post-acute sequelae of COVID-19

EHR: electronic health record

SARS-CoV-2: severe acute respiratory syndrome coronavirus 2

GO: Gene Ontology

PPI: protein-protein interaction

RAT: rapid antigen test

PCR: polymerase chain reaction

HKHA: Hong Kong Hospital Authority

SMD: standardized mean difference

RR: Relative Risk

REFERENCES

1. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133–146 (2023).
2. Zhang, H. *et al.* Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat. Med.* **29**, 226–235 (2023).
3. Russell, C. D., Lone, N. I. & Baillie, J. K. Comorbidities, multimorbidity and COVID-19. *Nat. Med.* **29**, 334–343 (2023).
4. Altmann, D. M., Whettlock, E. M., Liu, S., Arachchillage, D. J. & Boyton, R. J. The immunology of long COVID. *Nat. Rev. Immunol.* **23**, 618–634 (2023).
5. Xie, Y., Xu, E., Bowe, B. & Al-Aly, Z. Long-term cardiovascular outcomes of COVID-19. *Nat. Med.* **28**, 583–590 (2022).
6. Raman, B., Bluemke, D. A., Lüscher, T. F. & Neubauer, S. Long COVID: post-acute sequelae of COVID-19 with a cardiovascular focus. *Eur. Heart J.* **43**, 1157–1172 (2022).
7. Xie, Y., Xu, E. & Al-Aly, Z. Risks of mental health outcomes in people with covid-19: cohort study. *BMJ* **376**, e068993 (2022).
8. Xie, Y. & Al-Aly, Z. Risks and burdens of incident diabetes in long COVID: a cohort study. *Lancet Diabetes Endocrinol.* **10**, 311–321 (2022).
9. Bowe, B., Xie, Y., Xu, E. & Al-Aly, Z. Kidney Outcomes in Long COVID. *J. Am. Soc. Nephrol.* **32**, 2851 (2021).
10. Su, Y. *et al.* Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell* **185**, 881-895.e20 (2022).

Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing

11. Brodin, P. *et al.* Studying severe long COVID to understand post-infectious disorders beyond COVID-19. *Nat. Med.* **28**, 879–882 (2022).
12. Mehandru, S. & Merad, M. Pathological sequelae of long-haul COVID. *Nat. Immunol.* **23**, 194–202 (2022).
13. Lai, F. T. T. *et al.* Multimorbidity and adverse events of special interest associated with Covid-19 vaccines in Hong Kong. *Nat. Commun.* **13**, 411 (2022).
14. Kuan, V. *et al.* Identifying and visualising multimorbidity and comorbidity patterns in patients in the English National Health Service: a population-based study. *Lancet Digit. Health* **5**, e16–e27 (2023).
15. Subramanian, A. *et al.* Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat. Med.* **28**, 1706–1714 (2022).
16. Ayoubkhani, D. *et al.* Trajectory of long covid symptoms after covid-19 vaccination: community based cohort study. *BMJ* **377**, e069676 (2022).
17. Yang, J., Xu, Z., Wu, W. K. K., Chu, Q. & Zhang, Q. GraphSynergy: a network-inspired deep learning model for anticancer drug combination prediction. *J. Am. Med. Inform. Assoc.* **28**, 2336–2345 (2021).
18. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
19. ROSENBAUM, P. R. & RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).

20. Gomez-Cabrero, D. *et al.* From comorbidities of chronic obstructive pulmonary disease to identification of shared molecular mechanisms by data integration. *BMC Bioinformatics* **17**, 441 (2016).
21. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl. Acad. Sci.* **118**, e2025581118 (2021).
22. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
23. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
24. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
25. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
26. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
27. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
28. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).

Chaos
An Interdisciplinary Journal
of Nonlinear Science

AIP Publishing

29. Yu, G. Gene Ontology Semantic Similarity Analysis Using GOSemSim. in *Stem Cell Transcriptional Networks: Methods and Protocols* (ed. Kidder, B. L.) 207–215 (Springer US, New York, NY, 2020). doi:10.1007/978-1-0716-0301-7_11.

30. Rivals, I., Personnaz, L., Taing, L. & Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**, 401–407 (2007).

31. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

32. Ramos, J. & others. Using tf-idf to determine word relevance in document queries. in *Proceedings of the first instructional conference on machine learning* vol. 242 29–48 (Citeseer, 2003).

33. Xu, Z., Zhang, J., Zhang, Q., Xuan, Q. & Yip, P. S. F. A comorbidity knowledge-aware model for disease prognostic prediction. *IEEE Trans. Cybern.* **52**, 9809–9819 (2021).

34. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).

35. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

36. Borczuk, A. C. & Yantiss, R. K. The pathogenesis of coronavirus-19 disease. *J. Biomed. Sci.* **29**, 87 (2022).

Chaos
An Interdisciplinary Journal
of Nonlinear Science

597 37. van Doorn, A. S., Meijer, B., Frampton, C. M. A., Barclay, M. L. & de Boer, N. K. H.
598 Systematic review with meta-analysis: SARS-CoV-2 stool testing and the potential for
599 faecal-oral transmission. *Aliment. Pharmacol. Ther.* **52**, 1276–1288 (2020).
600 38. Xiao, F. *et al.* Evidence for Gastrointestinal Infection of SARS-CoV-2. *Gastroenterology*
601 **158**, 1831-1833.e3 (2020).
602 39. Zhou, Y. *et al.* Network medicine links SARS-CoV-2/COVID-19 infection to brain
603 microvascular injury and neuroinflammation in dementia-like cognitive impairment.
604 *Alzheimers Res. Ther.* **13**, 110 (2021).
605 40. Liang, Y. *et al.* γ-Linolenic Acid Prevents Lipid Metabolism Disorder in Palmitic Acid-
606 Treated Alpha Mouse Liver-12 Cells by Balancing Autophagy and Apoptosis via the
607 LKB1-AMPK-mTOR Pathway. *J. Agric. Food Chem.* **69**, 8257–8267 (2021).
608 41. Yang, D. *et al.* Targeting intracellular Neu1 for coronavirus infection treatment. *iScience*
609 **26**, 106037 (2023).
610 42. Satu, M. S. *et al.* Diseasome and comorbidities complexities of SARS-CoV-2 infection
611 with common malignant diseases. *Brief. Bioinform.* **22**, 1415–1429 (2021).
612 43. White, E. J. *et al.* Sialidase down-regulation reduces non-HDL cholesterol, inhibits
613 leukocyte transmigration, and attenuates atherosclerosis in ApoE knockout mice. *J. Biol.*
614 *Chem.* **293**, 14689–14706 (2018).
615 44. Zhang, C., Chen, J., Liu, Y. & Xu, D. Sialic acid metabolism as a potential therapeutic
616 target of atherosclerosis. *Lipids Health Dis.* **18**, 173 (2019).

45. Jensen-Cody, S. O. & Potthoff, M. J. Hepatokines and metabolism: Deciphering communication from the liver. *Mol. Metab.* **44**, 101138 (2021).

46. Sekiyama, K., Ushiro, Y., Kurisaki, A., Funaba, M. & Hashimoto, O. Activin E enhances insulin sensitivity and thermogenesis by activating brown/beige adipocytes. *J. Vet. Med. Sci.* **81**, 646–652 (2019).

47. Gu, J. *et al.* The role of lysosomal membrane proteins in glucose and lipid metabolism. *FASEB J.* **35**, e21848 (2021).

48. Sleat, D. E. *et al.* Genetic evidence for nonredundant functional cooperativity between NPC1 and NPC2 in lipid transport. *Proc. Natl. Acad. Sci.* **101**, 5886–5891 (2004).

49. Benjamin, D. I. *et al.* Ether lipid generating enzyme AGPS alters the balance of structural and signaling lipids to fuel cancer pathogenicity. *Proc. Natl. Acad. Sci.* **110**, 14912–14917 (2013).

50. Ogburn, E. L., Sofrygin, O., Díaz, I. & van der Laan, M. J. Causal Inference for Social Network Data. *J. Am. Stat. Assoc.* **119**, 597–611 (2024).

51. Bang, S., Kim, J.-H. & Shin, H. Causality modeling for directed disease network. *Bioinformatics* **32**, i437–i444 (2016).

52. Feuerriegel, S. *et al.* Causal machine learning for predicting treatment outcomes. *Nat. Med.* **30**, 958–968 (2024).

53. Belyaeva, A., Squires, C. & Uhler, C. DCI: learning causal differences between gene regulatory networks. *Bioinformatics* **37**, 3067–3069 (2021).