RAMEN identifies effective indicators for severe COVID and Long COVID patients

Yiwei Xiong[1†], Jingtao Wang[1†], Xiaoxiao Shang[1], Tingting Chen[1], Gregory Fonseca[1], Simon Rousseau[1*], and Jun Ding[1*]

[1]Meakins-Christie Laboratories, Research Institute of McGill Health Centre, Montreal, QC, Canada

[†] All those authors contributed equally. [*] Correspondence should be addressed to SR (simon.rousseau@mcgill.ca) and JD (jun.ding@mcgill.ca)

Abstract

The outbreak of the COVID-19 pandemic caused catastrophic socioeconomic consequences and fundamentally reshaped the lives of billions across the globe. Our current understanding of the relationships between clinical variables (demographics, symptoms, follow-up symptoms, comorbidities, treatments, lab results, complications, and other clinical measurements) and COVID-19 outcomes remains obscure. Various computational approaches have been employed to elucidate the relationships between different COVID-19 clinical variables and their contributions to the disease outcomes. However, it is often challenging to capture the indirect relationships, as well as the direction of those relationships, with the conventional pairwise correlation methods. Graphical models (e.g., Bayesian networks) can address these limitations but are computationally expensive, which substantially limits their applications in reconstructing relationship networks of umpteen clinical variables. In this study, we have developed a method named RAMEN, which employs Genetic Algorithm and random walks to infer the Bayesian relationship network between clinical variables. We applied RAMEN to a comprehensive COVID-19 dataset, Biobanque Québécoise de la COVID-19 (BQC19). Most of the clinical variables in our reconstructed Bayesian network associated with COVID-19 severity, or long COVID, are supported by existing literature. We further computationally verified the effectiveness of the RAMEN method with statistical examinations of the multi-omics measurements (Clinical variables, RNA-seq, and Somascan) of the BQC19 data and simulations. The accurate inference of the relationships between clinical variables and disease outcomes powered by RAMEN will significantly advance the development of effective and early diagnostics of severe COVID-19 and long COVID, which can help save millions of lives.

Keywords: Post-COVID Conditions, Long COVID, COVID-19, COVID-19 Severity, Genetic Algorithm, Random Walk, Bayesian Network, Clinical Variables, Multi-omics

INTRODUCTION

The outbreak of the COVID-19 pandemic has reshaped billions of lives across the globe and caused catastrophic socioeconomic losses in the past years[1,2]. Despite the excessive damage that COVID-19 has brought to humanity, our understanding of COVID-19 disease mechanisms remains elusive, which substantially limits the diagnostics and therapeutics for COVID-19. For instance, it is known that different COVID-19 patients may develop distinct symptoms and clinical outcomes[3]. However, it remains unclear why some people tend to develop more severe COVID-19 outcomes while some barely show any symptoms. This lack of understanding restricts the early clinical diagnosis and interventions for the most vulnerable COVID-19 victims[4,5], which leads to undo suffering. Moreover, although most COVID-19 victims can recover from the disease, approximately 30% of them develop long-term symptoms (termed as long COVID or post-COVID-19 conditions)[4], which has a tremendous socioeconomic impact. Long COVID patients suffer from physical and psychological pain and discomfort[5]. Many have

lost their ability to work, significantly contributing to the spike of ongoing inflation in the past few years that are detrimental to billions of people's quality of life. The fast-accumulating COVID-19 datasets present unprecedented opportunities to derive a deeper understanding of the disease mechanisms underlying COVID-19, which will drive the discovery of novel diagnostic markers and therapeutic targets. Multiple initiatives across the globe are profiling clinical variables and genomics sequencing associated with COVID-19 patients[6,7]. In Quebec, the Biobanque québécoise de la COVID-19 (BQC19)[8] has collected clinical data for over three thousand COVID-19 patients, along with data of several other modalities.

With the ever-increasing availability of COVID-19 datasets, many studies interrogated the complex relationships between various clinical variables and COVID-19 severities[9-12] using statistical methods (e.g., correlation[12]) and machine learning methods (e.g., regression tree[9]). Many existing methods can well capture the direct relationships between clinical variables and clinical outcomes (edges in the network). Still, the category of correlation-based approaches can only infer the association between variables, and the direction of these relationships remains unknown. In addition, although the correlation-based approaches may identify direct relationships between two variables, those methods often fail to capture indirect ones, particularly in a complex network involving hundreds of variables. Furthermore, correlation-based methods focus on the pairwise relationship between pairs of variables, and the overall relationships among multiple (more than 2) variables are challenging to identify. On the other hand, Bayesian Network (BN)[13-15], a probabilistic graphical model, can somehow address the above limitations of correlation-based approaches, enabling the inference of clinical variables that could indicate COVID-19 outcomes (e.g., COVID-19 severity or long COVID). Previously, BNs have shown success in many application scenarios and outperformed physicians in disease diagnostics[16,17]. For example, a BN-based diagnosis not only achieved state-of-the-art overall performance for neurogenerative diseases compared to a list of other methods[18-23] but also provided very good interpretability. Nevertheless, building a Bayesian network (particularly structure learning) for hundreds and thousands of clinical variables is challenging and computationally expensive. This task is complicated by the huge search space of possible solutions (In fact, the problem is NP-hard[24]). In practice, the Bayesian network structure learning is often regularized by prior knowledge (e.g., known constraints) to reduce the search space[25,26]. Unfortunately, this is not an option for COVID-19 since our current understanding of this relatively new disease remains very limited[27]. Besides, we are also interested in how each clinical variable influences COVID-19 outcomes (COVID-19 severity or long COVID).

Because of the above limitations, no existing methods can be directly applied to infer the complex relationships between clinical variables in COVID-19. Therefore, here we propose RAMEN (**R**andom w**a**lk and genetic Algorith**m** based n**e**twork infere**n**ce) that glues random walk and Genetic Algorithm to infer a Bayesian network that represents the relationships between clinical variables in COVID-19 (particularly between clinical variables and COVID-19 outcomes). The random walk is employed to rank and select the most relevant variables to COVID-19 outcomes to reduce the network complexity. The Genetic Algorithm will then be leveraged to find an optimized network structure that best fits the observations. To examine the performance of RAMEN, we have applied the method on two different COVID-19 cohorts from the BQC19 project with multi-omics measurements. RAMEN demonstrated superior performance in reconstructing the relationships between clinical variables of COVID-19. Many variables we predicted were previously reported to be associated with COVID-19 severity confirming the validity of RAMEN as a predictive and interpretable tool. RNA-seq and

Somascan data from the same COVID-19 cohorts also verified those relationships. Next, RAMEN identified novel clinical variables strongly linked to long COVID, for which our knowledge remains minimal. For example, early joint pain symptoms (within the first month of COVID-19 infection) strongly indicate the risk of long COVID. The capability of our model in predicting strong early indicators for COVID-19 outcomes (severity and long COVID) will significantly enhance the development of effective early diagnoses of patients who will develop severe COVID-19 symptoms or long COVID, which will help physicians to make personalized diagnoses and interventions that will save millions of lives.

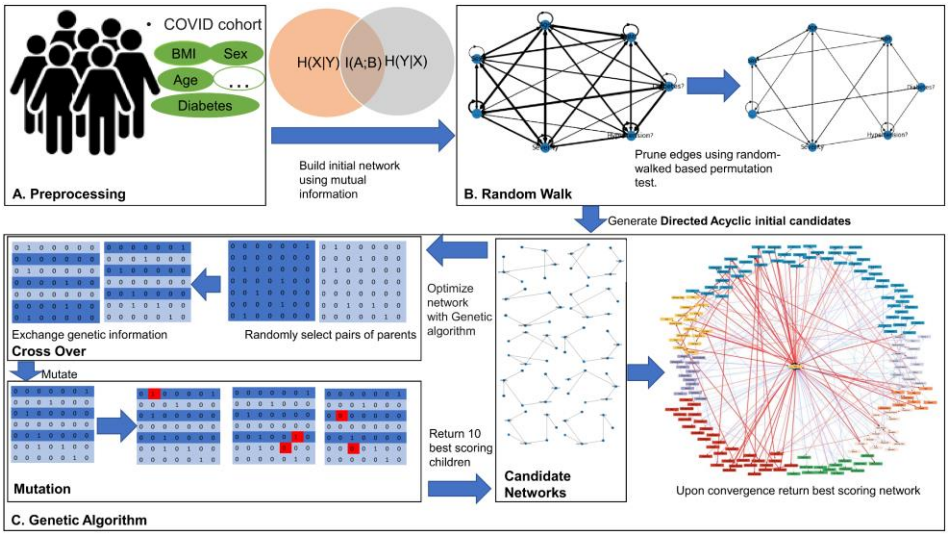

Figure 1 RAMEN overview. Given the clinical variable data of patients, our method learns the Bayesian Network with the following three steps: (1) Data Preprocessing, including removing variables that are not relevant to biology and variables with too many missing values; (2) Filtering irrelevant variables and getting an initial Bayesian network skeleton using a random-walk-based permutation test; (3) Using Genetic Algorithm to further search for the optimized network. Supplementary table S15 gives the full name of each

RESULTS

Overview of RAMEN

The Ramen method is composed of two major components: random walks to select the most relevant variables to the COVID-19 outcomes (severity or long COVID) and build a draft candidate network. Next, employ Genetic Algorithm to find the optimized network that represents the relationships between different clinical variables based on the draft network from the random walks (Fig.1). Random walks: First, we calculate mutual information[28] between every possible pair of clinical variables (all possible edges). All the calculated mutual information will be normalized and represented as the transition probability between clinical variables (edges). We then perform random walks starting from each non-terminal variable (all variables other than the COVID outcomes) for *N* random steps. A random walk will stop once it reaches the destination (absorbing terminal nodes: severity or long COVID) or runs out of steps. After the random walks, we count the number of visits for each of the edges in all successful random walks (the walks that reach the terminal destination within *N* steps). Next, we perform random permutations of the transition probabilities between all nodes (edges). With the random transition probabilities, we re-perform random walks to get the number of visits for all edges by random. Third, we then employ random permutations to filter edges that are not significantly visited. Genetic Algorithm: The network with all remaining significant edges (and the nodes that are connected by those edges) is used as the starting point to search for the Bayesian network with the Genetic Algorithm. First, we generate candidate parent networks from the initial network obtained with random walks. Next, we crossover all those parent networks to produce offspring. Third, each of the offspring networks mutate to produce more candidate networks. Fourth, all these candidate networks (parents, offspring, and their mutations) are scored to select the best networks as the parents for the next generation. We iterate the above 'evolution' process until convergence to obtain the final relationship network.

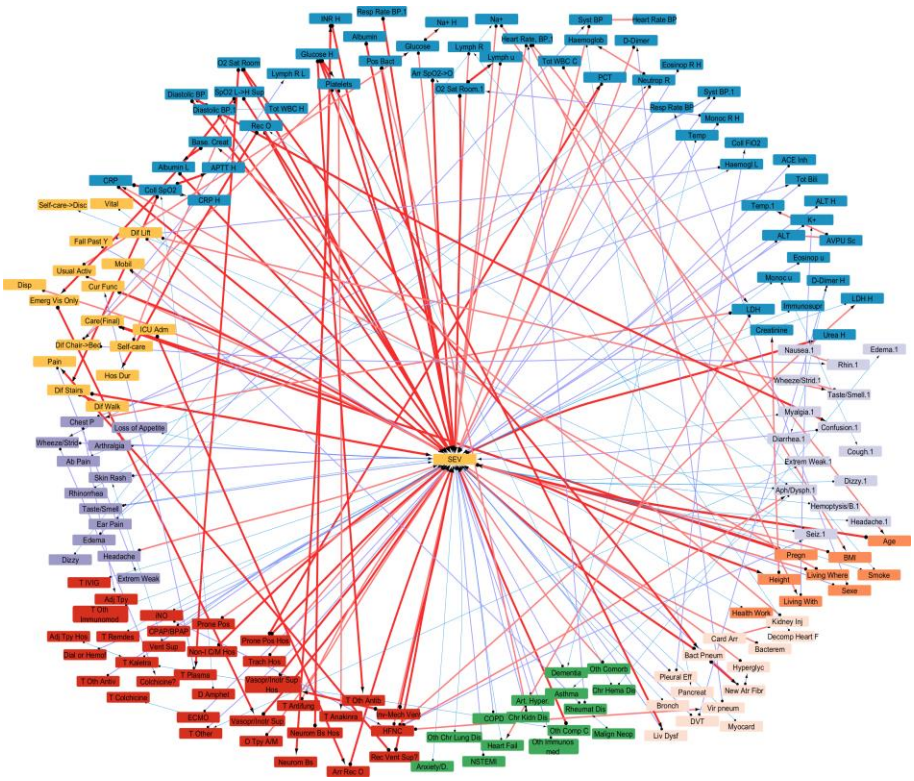

Figure 2: A compact COVID-19 severity network. The colors and thickness of the edges represent the strength: blue (low) to red (high), which is quantified by the rank of the edge mutual information. The node color represents variables of different categories (red: Treatments, purple: Symptoms, green: Comorbidity, blue: Lab results, orange: Demographics, pink: Complications, gray: follow-up symptom)

RAMEN reconstructs the COVID-19 severity network that describes significant relationships associated with COVID-19 severity

We employed the RAMEN method to the hospitalized patient cohort data from BQC19[8] and obtained a Bayesian network describing the complex relationships between clinical variables and their links to COVID-19 severity (Fig.2). The COVID-19 severity variable classifies patients' most severe condition into five levels: not infected, mild, moderate, severe, and deceased. The reconstructed relationship networks are well-supported by existing literature. For example, the top clinical variables linked to COVID-19 severity include Sex[10,29], Age[10,30], BMI[31,32], Arterial Hypertension[33], ALT[34], C-reactive protein (CRP) (highest value)[35], and Albumin (lowest value)[36]. To showcase the specificity of our method, we also simulated a dummy variable whose values are randomly generated using a uniform distribution. We treated this dummy variable as a simulated clinical variable and added it to the original BQC19 severity data (hospitalized patient cohort). This dummy variable was removed by the random walks of our approach, and RAMEN still produced the same Bayesian network as the original data, which demonstrated the specificity and robustness of our RAMEN method.

The COVID-19 severity network is supported by the COVID-19 hospitalized cohort

Heatmaps in Fig.3 show the percentage of patients with different severities for a

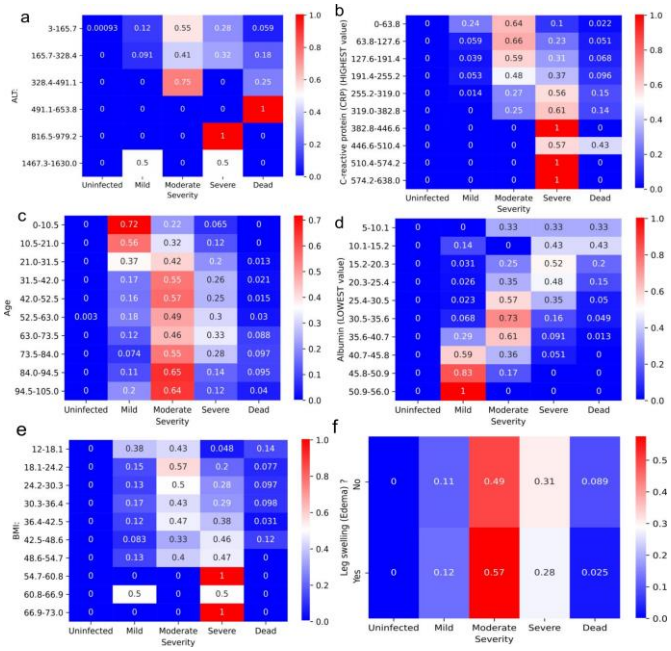

Figure 3: Heatmaps showing the relationships between clinical variables and COVID severity. Values in each entry are corresponding patient numbers normalized over rows (the sum of each row=1.0). Each row represents % of patients with different severity for the given value of the variable of interest.

different value range of clinical variables. Namely, Fig.3 heatmaps indicate whether one clinical variable could suggest specific COVID-19 severities. Here, we showed the heatmap for 5 top variables connected to COVID-19 severity-relevant variables (ALT, CRP-Highest value, Age, Albumin, BMI). In addition, to demonstrate the specificity of RAMEN, we also show the heatmap for a clinical variable not linked to COVID-19 severity (Leg swelling (Edema)). In general, for the positive variables that are directly linked to COVID-19 severity in our reconstructed network, the distribution of COVID-19 severity (% of patients across different severities) changes substantially accordingly to the value change of the relevant variables. On the other hand, the distribution of COVID-19 severity stays unchanged in response to the change of the irrelevant variable. See supplementary results for an extended version of the interpretation.

The COVID severity network constructed by RAMEN is supported by multi-omics data

We leveraged the BQC19 multi-omics data to verify the relationships (edges) in our reconstructed network. For the top COVID-19 indicator variables (clinical variables linked to COVID-19 severity given by our method), we have examined the gene expression (RNA-seq) and protein level (Somascan data) from patient groups with different values of those indicators. We performed differentially expressed (DE) gene/protein analysis for each variable (see details in the methods section). Heatmaps in Fig.4 present the DE genes (identified from the RNA-seq data) associated with a list of top indicators and COVID-19 severity. A similar set of heatmaps that we identified from the Somascan data are available in Supplementary Fig.S3. First, we found that the DE genes associated with the top indicators (Albumin, ALT, CRP, BMI), as predicted in our network, significantly overlap with the DE genes we identified for COVID-19 severity, which verifies the relationships between those predicted clinical variables and the

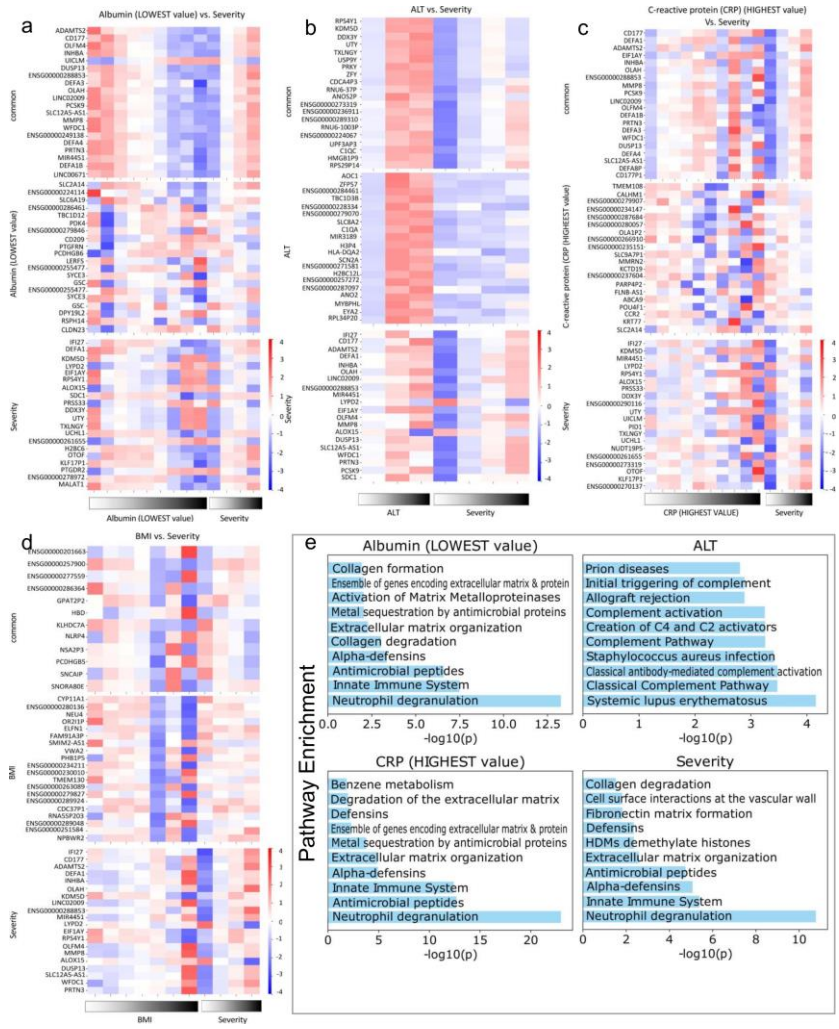

Figure 4: Examining edges in the reconstructed COVID severity network with the multi-omics data. Panel (a) shows the gene expression for three groups of differential genes associated with Albumin. We first identify all differential genes associated with Albumin (i.e., genes whose expression changes among the patients with different levels of Albumin (G_1). We then identify the differential genes associated with COVID severity (G_2). 'Common' shows the shared genes between those two groups of genes ($G_1 \cap G_2$); 'Albumin' shows the unique differential genes for the Albumin variable ($G_1 \cap \neg G_2$); 'Severity' shows the unique differential genes for COVID severity ($G_2 \cap \neg G_1$). Heatmaps in panels (b-d) have similar meanings as (a). Panel e shows the top enriched pathways for each of the variables based on their differential genes.

severity of the disease. More importantly, the multi-omics data also presents an unrivaled opportunity to understand the biological mechanism underlying those relationships. Specifically, we have performed pathway enrichment analysis for the DE genes that we identified for COVID-19 severity and its top indicator variables. Enriched Pathways like "Neutrophil degradation," "Innate Immune System," "Collagen Degradation," "Extracellular matrix organizations," and "Alpha defensins" are generally shared between DE genes for severity and DE genes for other clinical variables, which suggests the potential roles of those enriched pathways in modulating COVID-19 severity and how the indicator variable may influence the COVID-19 severity via those shared pathways. To our best knowledge, this is one of the very first, if not the first, methods that leverage multi-omics to systematically examine the biological mechanisms underlying the relationship between COVID-19 severity and hundreds of clinical variables (Fig.4).

RAMEN reconstructs the Long COVID relationship network

We next applied RAMEN to the outpatient COVID-19 cohort to study long COVID. Whether or not a patient has long COVID is classified according to a commonly used criterion: at least one symptom in three months[37]. In the BCQ-19 outpatient cohort for studying long COVID, long COVID frequency is 36.5% (526 out of 1440), close to the previously reported value[4]. One of our objectives is to identify critical indicators for long COVID that may assist in the early diagnosis. In this long COVID dataset, we only used clinical variables profiled within one month after COVID infection. All clinical variables curated after the first month were ignored in the analysis. The method reconstructed a Bayesian network for long COVID with dozens of clinical variables that are strongly linked to long COVID (Fig.5). Many known critical variables for long COVID, such as Age[38], Chest pain[5] , Joint Pain[39] , Runny nose[39], and Shortness of breath[5] were captured by RAMEN . On the other hand, irrelevant variables for long COVID, such as chronic kidney disease, are not shown in the network. This variable, to our best knowledge, is not reported to indicate a higher risk of long COVID. As discussed in one recent study[40], an in-depth assessment of kidney outcomes in the post-acute phase of COVID-19 infection is not yet available. Moreover, we also simulated a new dataset with several random dummy variables (random binary and categorical variables). RAMEN is able to produce exactly the same results as the original data, which indicates the specificity and robustness of our method.

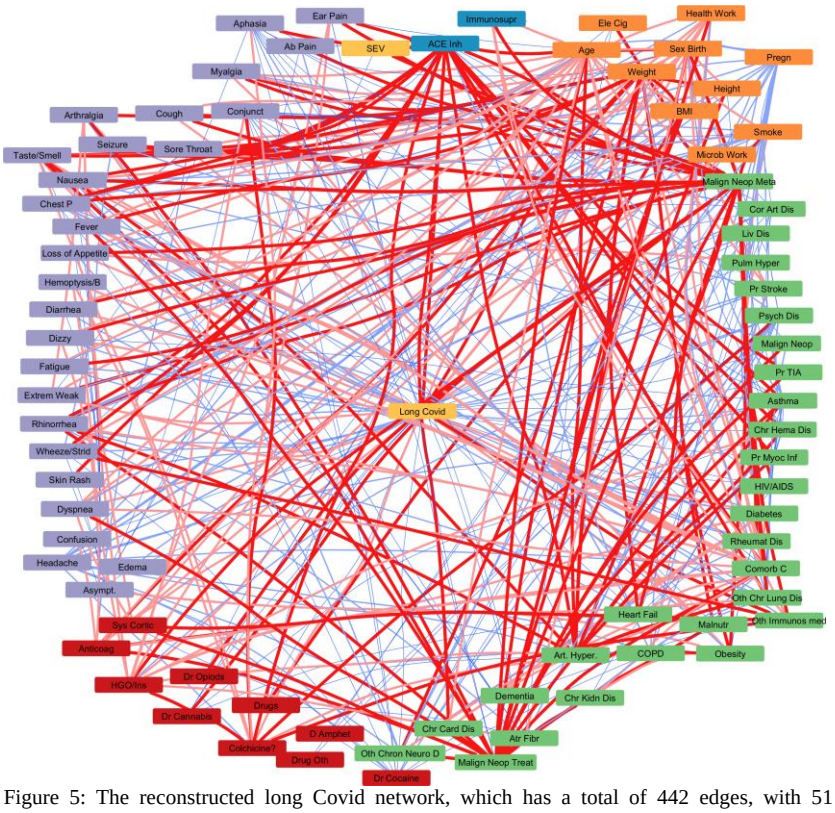

Figure 5: The reconstructed long Covid network, which has a total of 442 edges, with 51 connections to Long Covid. The colors and thickness of the edges represent the edge strength and the color of the nodes represents clinical variables of different categories.

The long covid network is supported by the BQC19 outpatient cohort

Chest pain, joint pain, Shortness of breath, and Age are the four top (highest mutual information) variables connected to long COVID in our network, and Fig. 6 shows that they have positive correlations with the long COVID variable. A higher value of any of these four variables corresponds to a higher risk of long COVID. The 'runny nose' variable is also connected to long COVID in our network, but it has a negative correlation with the long COVID. Namely, an early symptom of a 'runny nose' indicates a lower probability of developing long COVID. Please refer to the supplementary results for a more detailed discussion.

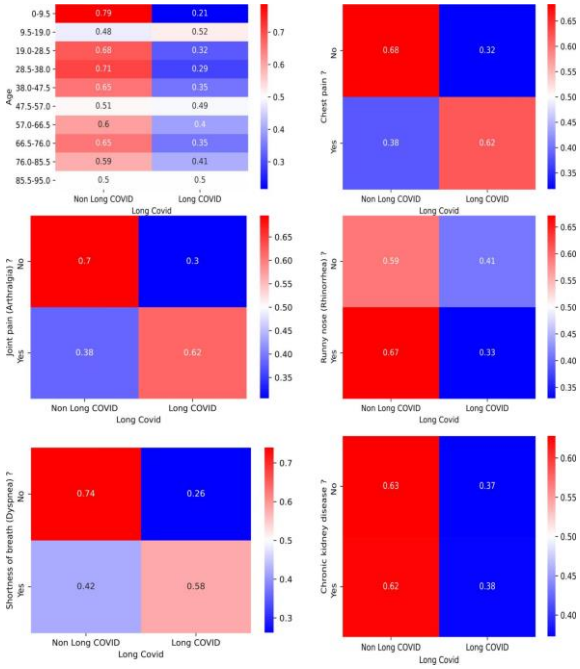

Figure 6: Heatmaps showing the relationships between clinical variables and long COVID. The meaning of the heatmaps is identical as in Fig. 3.

The long COVID network built by RAMEN is supported by COVID multi-omics data

Supplementary Fig.S7 and S10 show edge validations using BQC19 RNA-seq data and Somascan data, respectively. Long COVID's DE genes have significant overlap with DE genes for Shortness of breath, joint pain, and chest pain, suggesting potential common genes and biological pathways underlying those clinical variables and long COVID. We performed pathway enrichment analysis for the DE genes of each variable and presented our results in the bottom panels of Supplementary Fig.S7. Although common pathways exist (e.g., "Cell surface interactions at the vascular wall"), their connections to long COVID still need further examination.

Systematic long COVID network evaluation

We systematically examine the accuracy of RAMEN by using Pearson's chi-square test[41] to check the dependency between the long COVID variable and other clinical variables. The top 10 variables with the highest mutual information with long COVID all pass the test (see a list of variables and corresponding p values in Supplementary table S12), meaning the clinical variables and long COVID are dependent. To calculate an overall statistical significance, we repeated the above analysis for 10 random clinical variables in the network that are NOT connected to long COVID (variables that passed the random walk permutation test but were predicted to be not directly linked to long COVID). On average, there are only 1.78 (based on 100 random tests) random variables showing significant influences on long COVID patient distribution (most are independent of long COVID), which suggests that the predicted variables in our network are significantly supported by the clinical data (p-value=1-pbinom (10-1,10,0.178) =3.193e-08).

DISCUSSION

We have developed RAMEN, an efficient Bayesian network inference method for identifying complex relationships between clinical variables from clinical health records data of cohorts. The method employs random walks and Genetic Algorithm to efficiently search for the optimized Bayesian network structure that represents complex relationships between a large number of variables. We applied the method on two separate COVID-19 cohorts in the BQC19 dataset to reconstruct the networks for COVID severity (hospitalized COVID-19 cohort) and for long

COVID (COVID-19 cohort composed of outpatients with multiple revisits). As a result, RAMEN successfully captured critical indicator variables for COVID severity and long COVID, as supported by existing literature. In addition, those novel predicted relationships by RAMEN were further verified with the multi-omic measurements in BQC19, which demonstrates the effectiveness of the method.

Compared to other existing methods, RAMEN has three major unique features, which allow efficient reconstruction of a network that describes the complex relationships among hundreds of clinical variables in the disease cohort. First, RAMEN can reconstruct directed relationships (direct and indirect) between a myriad of clinical variables. The commonly used correlation-based methods often ignore the direction of the relationships between variables (please refer to supplementary results "Comparison with correlation-based approaches" section for details). However, such information is critical for understanding the complex relationships underlying diseases as it tells whether a certain variable is a "driver" or "effector" for the disease outcome of interest. On the other hand, existing Bayesian network inference methods are mostly very inefficient, which limits their application to a large network composed of hundreds of variables. By employing effective random walks and Genetic Algorithm, RAMEN can search for an optimized solution (a local optimum) very efficiently, enabling its scalable applications. In fact, we applied the method to the BQC19 data with 880 clinical variables, and we managed to get the converged results within 1 hour (Intel(R) Xeon(R) Platinum 8160 CPU @ 2.10GHz, 64G RAM). Second, RAMEN can enable the disease outcome of interest as the "terminal absorbing node" in the network, which essentially makes all the relationships in the network "conditional" on the disease outcome. In other words, the relationships between variables in the network (e.g., A→B) are quantified by the number of visits of B in all the successful random walks from A to the clinical outcome of interest (e.g., COVID severity), which is different from a direct correlation between the variables A and B. This characteristic could surely help to identify influential relationships for the disease outcome of interest, and such a feature is missing from most existing approaches. Third, RAMEN is one the very first methods/pipelines, if not the very first, that leverage multi-omics measurements to validate the inferred Bayesian network. The ever-increasing availability of multi-omics data actually offers opportunities to understand the biological mechanisms underlying the disease. Unfortunately, this has never been integrated into the existing methods/pipelines to provide systematic examinations of the reconstructed variable networks. Last, RAMEN provides an interactive web portal to allow users to interrogate the complex relationships between any variable of their interest against COVID outcomes.

Given the unique features discussed above, RAMEN presents an unprecedented toolset to explore the complicated relationships between clinical variables in complex diseases (such as COVID-19). This will shed light on the discovery of biomarkers (or indicators) for early and effective diagnosis of the disease, which could allow early interventions and special attention from the physicians to protect the patients from developing complications. For example, early joint pain is a strong indicator of long COVID risk. Therefore, patients who show early joint pain symptoms may demand early interventions to reduce the long COVID risk. Overall, we believe that RAMEN empowers efficient and accurate examinations of complex relationships between clinical variables (and with disease outcomes), which will significantly enhance the development of early and effective diagnosis and interventions for severe and long COVID, greatly reducing the socioeconomic costs of COVID-19 (and particularly long COVID). The tool is general and could be applied to study other disease cohorts with similar data structures.

METHODS

Datasets

In this study, we use the multi-modality data (clinical variables, RNA-seq, and Somascan) for two cohorts (hospitalized and outpatients) from BQC19. A significant portion of the patients also profiled their gene expressions (RNA-seq) and protein levels (Somascan). The clinical variables associated with COVID-19 patients include symptoms (e.g., cough, muscle pain), comorbidities (e.g., asthma, diabetes), demographics (e.g., sex, age), lab test results (e.g., white blood cell counts, ALT), and others. We use hospitalized patients to study the relationships between COVID-19 severity and other variables because hospitalized patients have few missing values in their clinical records, especially lab results. In comparison, outpatients have more follow-up visits, and thus the outpatient cohort is more suitable for studying long COVID.

Data Preprocessing

We processed the BQC19 clinical variable dataset with the following steps to make it suitable for subsequent models: (1) Combining visits and creating one record for each patient; (2) Removing variables with too many missing values (Supplementary Fig.S13 & S14 show the cutoff); (3) Discretizing real-value variables to simplify mutual information computation. Besides the clinical variables, RNA-seq data were processed following a standard RNA-seq pipeline. The reads were mapped to the human reference genome using HISAT2[42] after quality control. The mapped reads were then used to quantify raw gene expression count using HTSeq-count[43]. The raw gene expression was then normalized using DESeq2[44]. Similarly, the Somascan[45] data were processed and normalized using a standard pipeline. Please check Supplementary Methods for more details.

Build initial network structure with a random-walk-based permutation test

After preprocessing, the first step of the network structure learning process is using a permutation test based on random walks to get an initial network structure. The approach starts with a fully connected directed graph where nodes are clinical variables, and edge weights are mutual information between the connected nodes.

$$I(X,Y) = -\Sigma_x P(x)\log P\,(x) - \Sigma_y P(y)\log P\,(y) + \Sigma_x\Sigma_y P(x,y)\log\big(P(x,y)\big) \quad (1)$$

We employ random walks to obtain an initial fully connected network. The random walk starts at any node X and terminates whenever the terminal node (Severity or Long COVID) is reached, or a certain number of steps has been completed. A random walk is considered 'successful' if the terminal absorbing node is reached. At each node X, the transition probability from node X to another node Y is given by the normalized mutual information between X and Y, which is described in $eqn.(2)$.

$$P_X(Y) = \frac{I(X,Y)}{\Sigma_N I(X,N)} \quad (2)$$

Where N is every neighbor of X and itself. Intuitively, edges that are more relevant to the terminal absorbing node (COVID-19 severity or long COVID) will get more visits during all successful random walks (successful visits). To reduce the network inference complexity and obtain an initial network structure, we then employed a permutation test to calculate a significance p-value for each of the edges. A directed edge is regarded as significantly visited if the number of successful visits is significantly greater than the number of successful visits in random permuted networks. All the insignificant edges are removed from the network. If the edges of both directions between two variables are both significant, they are treated as special bidirectional edges, which is NOT allowed in the final Bayesian network and only one direction will be adopted.

Bayesian network inference using Genetic Algorithm

Genetic Algorithm, a score-based structure learning algorithm, is used to search the optimized directed Bayesian network, which captures the complex relationships between clinical variables based on the initial network backbone obtained from Random walks. As Bayesian network structure learning is an NP-hard problem, Genetic Algorithm was chosen here to efficiently search for a local optimum. The Genetic Algorithm starts with ten directed acyclic graph (DAG) candidates (parent networks), which are generated based on the initial network structure produced by the random walks. Then the following steps will be executed until convergence.

1.Crossover: Randomly pick two networks and exchange the elements in their adjacency graphs.
2.Mutation: For any network, randomly 1) add one or two consecutive edges; or 2) remove one or two consecutive edges; or 3) flip the direction of one or two consecutive edges.
3.Scoring: Score the original network candidates and new network generated by the step 1 and step 2 using the entropy-based scoring function with a network complexity regularization.

$$Score(Net) = -\left(\sum_{X_i:\, Pa(X_i)=\emptyset} H_D(X_i) + \sum_{X_i:Pa(X_i)\neq\emptyset} \left(H_D\left(\{X_i\}\, U\, Pa(X_i)\right) - \left(H_D\left(Pa(X_i)\right)\right)\right)\right) - \alpha|E| \quad (3)$$

Where H_D represents the entropy of a random variable. Pa gives the parents of a variable. $|E|$ is network complexity regularization and equals the number of edges. α is a hyperparameter controlling the weights of the corresponding regularization factors. This formula is derived from Kullback-Leibler (KL) divergence between the actual data likelihood, and the data likelihood factorized according to the Bayesian Network model[46]. The Genetic Algorithm aims to minimize such KL distance so that the data is well represented by the model.

Computational validation of reconstructed network with the multi-omics data

We employed different strategies to validate the relationships in the reconstructed network.
(1)Validation with records of clinical variables. To investigate the relationships between any clinical variables and COVID-19 outcome (COVID severity and long COVID), we construct contingency tables specifying patient numbers corresponding to pairs of variable values and employ Pearson's Chi-square test[41] to calculate a p-value to indicate if the two variables are dependent. (2)Validation using data from other modalities (gene expression and protein level). We first grouped patients based on the values of the variables and identified the differentially expressed genes among those patient groups corresponding to different values using Wilcoxon test[47] (corrected p-value <0.05) and |log2 fold change|>0.6[48]. Next, we can interrogate the potential biological functions associated with the link between the indicator and the target node (COVID-19 severity or long COVID) by examining the list of genes in the above two lists. Specifically, for each DE gene list above, we used Toppgene[49] to find enriched pathways associated with those clinical variables, which could help us better understand the biological mechanism. Next, we performed a hypergeometric test to examine whether there is a significant overlap between the two DE gene list (DE indicator vs. DE target), suggesting how potentially the indicator variable influence the target variable.

SOFTWARE AVAILABILITY

The RAMEN program, with source codes and documentation, is publicly available at https://github.com/mcgilldinglab/RAMEN. The reconstructed networks were visualized using Cytoscape[50]. In addition, we implemented a web portal (http://dinglab.rimuhc.ca/pgm/) that can interactively visualize the reconstructed networks from RAMEN.

ACKNOWLEDGEMENTS

This work was partially supported by CIHR PJT-180505, FRQS 295298, 295299, and NSERC RGPIN2022-04399 to JD. We thank the BQC19 initiative for granting us access to the multi-omic COVID data.

REFERENCES

1 Mofijur, M. *et al.* Impact of COVID-19 on the social, economic, environmental and energy domains: Lessons learnt from a global pandemic. *Sustainable production and consumption* **26**, 343-359 (2021).

2 Nicola, M. *et al.* The implications of the coronavirus pandemic (COVID-19): A review. *International journal of surgery* **78**, 185-193 (2020).

3 Ding, J. *et al.* A network-informed analysis of SARS-CoV-2 and hemophagocytic lymphohistiocytosis genes' interactions points to Neutrophil extracellular traps as mediators of thrombosis in COVID-19. *PLoS Computational Biology* **17**, e1008810 (2021).

4 Logue, J. K. *et al.* Sequelae in adults at 6 months after COVID-19 infection. *JAMA network open* **4**, e210830-e210830 (2021).

5 Raveendran, A., Jayadevan, R. & Sashidharan, S. Long COVID: an overview. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* **15**, 869-875 (2021).

6 Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **49**, D10 (2021).

7 Zuo, X., Chen, Y., Ohno-Machado, L. & Xu, H. How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles. *Briefings in Bioinformatics* **22**, 800-811 (2021).

8 Tremblay, K. *et al.* The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. *PloS one* **16**, e0245031 (2021).

9 Vrotsou, K., Rotaeche, R., Mateo-Abad, M., Machón, M. & Vergara, I. Variables associated with COVID-19 severity: an observational study of non-paediatric confirmed cases from the general population of the Basque Country, Spain. *BMJ open* **11**, e049066 (2021).

10 Li, X. *et al.* Clinical determinants of the severity of COVID-19: A systematic review and meta-analysis. *PLoS One* **16**, e0250602 (2021).

11 Torres-Ruiz, J. *et al.* Redefining COVID-19 severity and prognosis: the role of clinical and immunobiotypes. *Frontiers in immunology*, 3695 (2021).

12 Ghayda, R. A. *et al.* Correlations of clinical and laboratory characteristics of COVID-19: a systematic review and meta-analysis. *International journal of environmental research and public health* **17**, 5026 (2020).

13 Pearl, J. in *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA.* 15-17.

14 Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. (Morgan kaufmann, 1988).

15 Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques*. (MIT press, 2009).

16 Heckerman, D. E., Horvitz, E. J. & Nathwani, B. N. Toward normative expert systems: Part i the pathfinder project. *Methods of information in medicine* **31**, 90-105 (1992).

17 Heckerman, E. & Nathwani, N. Toward normative expert systems: part ii probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in medicine* **31**, 106-116 (1992).

18 John, G. H. & Langley, P. Estimating continuous distributions in Bayesian classifiers. *arXiv preprint arXiv:1302.4964* (2013).

19 Witten, I. H. & Frank, E. Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record* **31**, 76-77 (2002).

20 Le Cessie, S. & Van Houwelingen, J. C. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **41**, 191-201 (1992).

21 Kohavi, R. in *European conference on machine learning.* 174-189 (Springer).

22 Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* **55**, 119-139 (1997).

23 Quinlan, J. R. *C4. 5: programs for machine learning*. (Elsevier, 2014).

24 Scanagatta, M., Salmerón, A. & Stella, F. A survey on Bayesian network structure learning from data. *Progress in Artificial Intelligence* **8**, 425-439 (2019).

25 Agrahari, R. *et al.* Applications of Bayesian network models in predicting types of hematological malignancies. *Scientific reports* **8**, 1-12 (2018).
26 Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *Journal of computational biology* **7**, 601-620 (2000).
27 Bartoli, A., Gabrielli, F., Alicandro, T., Nascimbeni, F. & Andreone, P. COVID-19 treatment options: a difficult journey between failed attempts and experimental drugs. *Internal and emergency medicine* **16**, 281-308 (2021).
28 Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379-423 (1948).
29 Pradhan, A. & Olsson, P.-E. Sex differences in severity and mortality from COVID-19: are males more vulnerable? *Biology of sex Differences* **11**, 1-11 (2020).
30 Barek, M. A., Aziz, M. A. & Islam, M. S. Impact of age, sex, comorbidities and clinical symptoms on the severity of COVID-19 cases: a meta-analysis with 55 studies and 10014 cases. *Heliyon* **6**, e05684 (2020).
31 Caci, G. *et al.* COVID-19 and obesity: dangerous liaisons. *Journal of clinical medicine* **9**, 2511 (2020).
32 Chang, T. H., Chou, C. C. & Chang, L. Y. Effect of obesity and body mass index on coronavirus disease 2019 severity: a systematic review and meta-analysis. *Obesity Reviews* **21**, e13089 (2020).
33 Sisnieguez, C. E. L., Espeche, W. G. & Salazar, M. R. Arterial hypertension and the risk of severity and mortality of COVID-19. *European Respiratory Journal* **55** (2020).
34 Kumar-M, P. *et al.* Coronavirus disease (COVID-19) and the liver: a comprehensive systematic review and meta-analysis. *Hepatology international* **14**, 711-722 (2020).
35 Ali, N. Elevated level of C-reactive protein may be an early marker to predict risk for severity of COVID-19. *Journal of medical virology* (2020).
36 Paliogiannis, P. *et al.* Serum albumin concentrations are associated with disease severity and outcomes in coronavirus 19 disease (COVID-19): a systematic review and meta-analysis. *Clinical and Experimental Medicine* **21**, 343-354 (2021).
37 Greenhalgh, T., Knight, M., Buxton, M. & Husain, L. Management of post-acute covid-19 in primary care. *bmj* **370** (2020).
38 Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nature medicine* **27**, 626-631 (2021).
39 Aiyegbusi, O. L. *et al.* Symptoms, complications and management of long COVID: a review. *Journal of the Royal Society of Medicine* **114**, 428-442 (2021).
40 Bowe, B., Xie, Y., Xu, E. & Al-Aly, Z. Kidney outcomes in long COVID. *Journal of the American Society of Nephrology* **32**, 2851-2862 (2021).
41 Cressie, N. & Read, T. R. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)* **46**, 440-464 (1984).
42 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).
43 Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *bioinformatics* **31**, 166-169 (2015).
44 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).
45 Gold, L., Walker, J. J., Wilcox, S. K. & Williams, S. Advances in human proteomics at high scale with the SOMAscan proteomics platform. *New biotechnology* **29**, 543-549 (2012).
46 De Campos, L. M. & Friedman, N. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research* **7** (2006).
47 Wilcoxon, F. in *Breakthroughs in statistics* 196-202 (Springer, 1992).
48 Kang, H. C. *et al.* Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays. *Clinical Cancer Research* **10**, 272-284 (2004).
49 Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37**, W305-W311 (2009).
50 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).