Constructing Co-occurrence Network Embeddings to Assist Association Extraction for COVID-19 and Other Coronavirus Infectious Diseases

David Oniani,[1], Guoqian Jiang, M.D., Ph.D. [2], Hongfang Liu, Ph.D.[2], Feichen Shen, Ph.D.[2]

1. Kern Center for the Science of Health Care Delivery,
2. Division of Digital Health Sciences,
Mayo Clinic, Rochester MN

Corresponding author:

Feichen Shen, PhD
Division of Digital Health Sciences
Mayo Clinic
Rochester 55901 MN, USA
E-mail: shen.feichen@mayo.edu

Word count (up to 4000 words): 3,679
Structured Abstract (up to 250 words)**:** 250
Tables (up to 4)**:** 2
Figures (up to 6)**:** 3

Abstract

Objective

As COVID-19 started its rapid emergence and gradually transformed into an unprecedented pandemic, the need for having a knowledge repository for the disease became crucial. To address this issue, a new COVID-19 machine readable dataset known as COVID-19 Open Research Dataset (CORD-19) has been released. Based on this, our objective was to build a computable co-occurrence network embeddings to assist association detection amongst COVID-19 related biomedical entities.

Materials and Methods

Leveraging a Linked Data version of CORD-19 (i.e., CORD-19-on-FHIR), we first utilized SPARQL to extract co-occurrences among chemicals, diseases, genes, and mutations and build a co-occurrence network. We then trained the representation of the derived co-occurrence network using node2vec with four edge embeddings operations (L1, L2, Average, and Hadamard). Six algorithms (Decision Tree, Linear Regression, Support Vector Machine, Random Forest, Naive Bayes, and Multi-layer Perceptron) were applied to evaluate performance on link prediction. An unsupervised learning strategy was also developed incorporating the t-SNE and DBSCAN algorithms for case studies.

Results

Random Forest classifier showed the best performance on link prediction across different network embeddings. For edge embeddings generated using the Average operation, Random Forest achieved the optimal average precision of 0.97 and F1 score of 0.90. For unsupervised learning, 63 clusters were formed with silhouette score of 0.128. Significant associations were detected for five coronavirus infectious diseases in their corresponding subgroups.

Conclusion

In this study, we constructed COVID-19-centered co-occurrence network embeddings. Results indicated that the generated embeddings were able to extract significant associations for COVID-19 and coronavirus infectious diseases.

Keywords:

COVID-19; Coronavirus infectious diseases; Co-occurrence network embeddings; Association extraction

Introduction

Having now affected millions of people worldwide, COVID-19/Novel coronavirus has become a major pandemic of the century. Most of the countries have declared the state of national emergency and took actions effective immediately to slow the spread. Researchers and medical personnel around the world have published and released thousands of papers over a short period of time, covering a vast scientific ground and exploring medical treatments and possible vaccines for the virus [1]. With all this information, it is important to assemble all the available heterogeneous information and be aware of the explicit or implicit associations amongst subjects related to COVID-19 (e.g., certain genes could be linked to other genes and/or mutations related to COVID-19 and other coronavirus infectious diseases). Figuring out which subjects appear together is one of the approaches for identifying these associations and linking them together. Traditionally, text semantic similarity [2, 3] is one of the approaches for detecting links between words or sentences from unstructured data. One limitation is that it is inefficient to apply this approach over a large collection of free-text data, hampering the global view to detect significant associations across literature from heterogeneous domains. Normalized data stored in semi-structured graph format is more suitable for global link detection, as linked data [4] by nature provides efficient query scheme over triplets to interpolate between "macroscopic" and "microscopic" search.

Several efforts were made in the graph-based analysis of COVID-19. For example, Ahamed and Samad developed a graph-based model using abstracts of 10,683 COVID-19-related scientific articles and applying betweenness-centrality to rank order the importance of keywords related to drugs, diseases, pathogens, hosts of pathogens, and biomolecules [5]. Bellomarini et al. [6] present a report on ongoing work about the application of automated reasoning and knowledge graph technology to address the impact of the COVID-19 outbreak on the network of Italian companies. Tsiotas and Magafas used visibility graphs to study Greek COVID-19 infection-curve as a complex network [7]. Per request of the White House Office of Science and Technology Policy, new COVID-19 machine readable dataset (CORD-19) [8] has been released and several studies have featured the CORD-19 dataset to investigate COVID-19 related topics. For example, Wolinski has used CORD-19 for extract diseases at risk and calculate relevant indicators as well as created VIDAR-19 (VIsualization of Diseases At Risk in CORD-19) [9]. Wang, et al. have conducted CORD-19 named entity recognition leveraging the distant supervision strategy [10]. CORD-19-on-FHIR is a Linked Data version of CORD-19 [11]. It is represented in FHIR RDF [12, 13] and was produced by data mining the CORD-19 dataset and adding semantic annotations. In addition, Groza [14] has featured CORD-19-on-FHIR in the analysis of how semantically annotated dataset can be applied for detecting and preventing the potential spread of deceptive information regarding COVID-19.

A vast co-occurrence information contained in CORD-19 datasets allows for detection of novel associations across findings from various research articles. However, such information has been largely unexplored for association extraction. Moreover, the lack of measureable association amongst heterogeneous biomedical entities hampers the capability for a quantitative analysis. Inspired by the success of word embeddings [15] in building distributed semantic representations for each word given a corpus, network embeddings provide a solution to map graph nodes to distributional representations and translate nodes' relationships from graph space to embedding space, which makes the association between the nodes measurable [16-18]. In this study, we filled this gap by constructing network embeddings for the CORD-19 co-occurrence network. Specifically, we first derived a co-occurrence network by querying the CORD-19-on-FHIR and focused on the extraction of biomedical entities falling in four categories: chemical, disease, gene, and mutation. We then applied the node2vec model over the generated network and constructed network embeddings. We conducted the evaluation quantitatively and qualitatively. For the quantitative evaluation, we generated different embeddings with four embeddings generation operations using a downstream application on graph link prediction and measured the performance with six machine learning algorithms. For the qualitative evaluation, we visualized clusters generated by the optimal COVID-19 network embeddings and analyzed associations of heterogeneous biomedical entities related to COVID-19 and other coronavirus infectious diseases.

Materials

CORD-19-on-FHIR

The purpose of building CORD-19-on-FHIR is to represent linkage with other biomedical datasets and enable answering research questions. In this study, we used a subset of CORD-19-on-FHIR datasets annotated by Pubtator [19] and LitCovid [1], including 3,207 COVID-19 related articles in total. Each article was stored in one specific

annotation file. For each file and for each paragraph in the file, CORD-19-on-FHIR provides a way to capture all the annotated biomedical entities. A high level example of data stored in the Terse RDF Triple Language (Turtle) format is shown below.

```
pmc:annotations [
    pmc:id "1" ;
            pmc:infons [ pmc:identifier "MESH:D003371" ; pmc:type "Disease" ] ;
            pmc:locations [ pmc:length "5"^^xsd:int ; pmc:offset "20312"^^xsd:int ] ;
            pmc:text "cough" ],
    pmc:id "2" ;
            pmc:infons [ pmc:identifier "MESH:C000657245" ; pmc:type "Disease" ] ;
            pmc:locations [ pmc:length "19"^^xsd:int ; pmc:offset "14766"^^xsd:int ] ;
            pmc:text "2019-nCoV infection" ],],
 pmc:annotations [
    pmc:id "5" ;
            pmc:infons [ pmc:identifier "59272" ; pmc:ncbi_homologene "41448" ; pmc:type "Gene" ] ;
            pmc:locations [ pmc:length "31"^^xsd:int ; pmc:offset "1986"^^xsd:int ] ;
            pmc:text "angiotensin-converting enzyme 2" ],
    pmc:id "7" ;
            pmc:infons [ pmc:identifier "MESH:C000657245" ; pmc:type "Disease" ] ;
            pmc:locations [ pmc:length "19"^^xsd:int ; pmc:offset "14766"^^xsd:int ] ;
            pmc:text "2019-nCoV infection" ],],
```

where “pmc:annotations” was used to differentiate different paragraphs within a same article, “pmc:id” was used to indicate different biomedical entities along with entity type (“pmc:type”), location and offset (“pmc:location”), and the original text from literature (“pmc:text”). Such encoding of the data made it possible to easily detect co-occurrence of biomedical entities within a single paragraph for building the network across the literature.

Node2Vec

The node2vec model used a random walk-based sampling strategy to balance the graph homophily [20] and structural equivalence [21]. The reason we chose to use node2vec is its ability to learn node representations with a balance between the breadth first search (BFS) and depth first search (DFS), which is essential for learning associations in a graph with both local and global views.

Methods

The workflow of this study is made of three modules, including a CORD-19-on-FHIR-based co-occurrence network generation module, a network embeddings construction module, and an unsupervised learning module (Figure 1).

Figure 1. Study workflow.

Co-occurrence network generation

For each literature, we treated paragraph-level co-occurrence in this study. We first designed a SPARQL query statement to extract paragraph-level co-occurrence of biomedical entities from CORD-19-on-FHIR. Particularly, in order to largely collect coronavirus related diseases and comorbidities, we built a list of keywords for diseases and symptoms to constrain the searching space, which includes *COVID-19*, *SARS*, *pneumonia*, *fever*, *fibrosis*, *diarrhea*, *coronavirus*, *bronchitis*, *Ebola*, *influenza*, and *ZIKA*. We extracted co-occurrences between gene-disease, mutation-

disease and chemical-disease using the following SPARQL query by replacing “Biomedical_Entity” with “Gene”, “Mutation”, and “Chemical” respectively:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX fhir: <http://hl7.org/fhir/>
PREFIX pmc: <https://www.ncbi.nlm.nih.gov/pmc/articles#>
SELECT distinct ?pmc_id0 ?text0 ?pmc_id1 ?text1 (count(?text1) as ?count) WHERE {
?pmc pmc:annotations
[ pmc:id ?id0 ; pmc:text ?text0 ; pmc:infons
[ pmc:type ?type0 ; pmc:identifier ?pmc_id0 ] ] .
   FILTER ((?type0 = 'Disease') ) .
   {select * where{
                        ?pmc pmc:annotations
                        [ pmc:id ?id1 ; pmc:text ?text1 ; pmc:infons
                        [ pmc:type ?type1 ; pmc:identifier ?pmc_id1 ] ] .
        FILTER ((?type1=Biomedical_Entity) && (contains (lcase(str(?text1)), "coronavirus") || contains
(lcase(str(?text1)), "sars") || contains (lcase(str(?text1)), "covid-19") || contains (lcase(str(?text1)), "pneumonia") ||
contains (lcase(str(?text1)), "fever") || contains (lcase(str(?text1)), "fibrosis") || contains (lcase(str(?text1)),
"diarrhea") || contains (lcase(str(?text1)), "bronchitis") || contains (lcase(str(?text1)), "ebola") || contains
(lcase(str(?text1)), "influenze") || contains (lcase(str(?text1)), "zika"))).

     }
   }
}Group by ?pmc_id0 ?text0 ?pmc_id1 ?text1 Order by DESC(?count)
```

The outputs of the query were composed of a list of pairwise biomedical entities with co-occurrence frequency. We then built a network based on this list by adding a link between any two biomedical entities if they have co-occurred at least once. As shown in Figure 1, the co-occurrence network was represented by source-target pairs, which were then used as input data for training node representations.

Network embeddings representation learning

We applied the node2vec model in this module. Node2vec implements a 2nd order random walk over the graph topological structure, denoting that three types of node are involved in a specific walk, namely source entity, intermediate entity, and target entity. Given any source entity as E_s, target entity as E_t, intermediate entity that exists on the path between E_s and E_t as E_i, normalization constant as Z, the distribution of entity E_t with a fixed length of random walk can be represented as:

$$P(E_t|E_i) = \begin{cases} \frac{\pi(E_i, E_t)}{Z} & if\ (E_i,E_t)\ is\ an\ edge \\ 0 & otherwise \end{cases} \quad (1)$$

where $\pi(E_i, E_t)$ is a transition probability between entities E_i and E_t. Given the weight over edge (E_i,E_t) as $w(E_i,E_t) = 1$, $\pi(E_i, E_t)$ could be calculated as:

$$\pi(E_i, E_t) = \alpha(E_s,E_t) \cdot w(E_i,E_t) \quad (2)$$

$\alpha(E_s,E_t)$ is a searching bias term developed in node2vec. Specifically, node2vec introduced two hyperparameters p and q to balance between the BFS and DFS searching strategies for both local and global optimization. Given the shortest distance between E_s and E_t as $sd(E_s,E_t)$, α for entities E_s and E_t is computed based on p and q:

$$\alpha(E_s,E_t) = \begin{cases} \frac{1}{p} & if\ sd(E_s,E_t) = 0 \\ 1 & if\ sd(E_s,E_t) = 1 \\ \frac{1}{q} & if\ sd(E_s,E_t) = 2 \end{cases}$$

After learning the sampled network data using random walk, we then leveraged the Skip-gram model to train entity representations on the sampled data. For each entity node $E_s \in E$ and all its sampled neighbors $\mathrm{N}(E_s)$, the loss function for entity representation learning could be described as:

$$\max_f \sum_{E_s \in E} \log P(\mathrm{N}(E_s)|f(E_s)) \qquad (4)$$

In the end, we normalized the prediction distribution by using a nonlinearity (e.g., softmax) and optimize this loss function using Stochastic gradient descent.

Unsupervised clustering of network embeddings

To render the relatively high-dimensional embedding representations of network embeddings into a lower-dimensional space, we utilized the t-distributed stochastic neighbor embedding (t-SNE) algorithm [22] to render the embeddings for all entity nodes into a 2D space. t-SNE does not perform clustering in and of itself, but instead renders each node embedding into a (x,y) coordinate. As such, additional post-processing is needed to re-group these points into discrete clusters. The Density-based spatial clustering of applications with noise (DBSCAN) algorithm [23] was therefore used over output generated by the t-SNE to further partition different entity groups into distinct clusters. Given a parameter ε that denotes how close points should be to each other and another parameter k that indicates the minimum number of points, the DBSCAN clustered similar entity nodes together based on density according to the pre-defined two parameters.

Experiments

From CORD-19-on-FHIR, we extracted 49,696 co-occurred biomedical entities for 3,626 coronavirus related diseases, 5,741 genes, 524 mutations, and 6,878 chemicals. Thus the derived co-occurrence network contains 16,769 nodes and 49,696 edges in total.

For quantitative evaluation, we generated the optimal network embeddings by performing a downstream link prediction task. Link prediction is a procedure where the goal is to predict the relationship between any two nodes and use the performance of a prediction to evaluate the quality of the generated network embeddings. Edge embeddings were used in this task in order to investigate the relationships between nodes leveraging distributional representations provided by entity embeddings. For any given nodes E_s, E_t and their corresponding entity representations $f(E_s)$ and $f(E_t)$, edge embeddings were calculated using four operations, namely Hadamard, Average, L1 and L2 as shown in Equations 5-8 respectively:

$$Hadamard(E_s, E_t) = f(E_s) * f(E_t) \qquad (5)$$

$$Average(E_s, E_t) = \frac{f(E_s) + f(E_t)}{2} \qquad (6)$$

$$L1(E_s, E_t) = |f(E_s) - f(E_t)| \qquad (7)$$

$$L2(E_s, E_t) = |f(E_s) - f(E_t)|^2 \qquad (8)$$

We used six conventional classification algorithms to evaluate the performance of different edge embeddings on link prediction task, including Decision Tree (DT) [24], Linear Regression (LR) [25], Support Vector Machine (SVM) [26], Random Forest (RF) [27], Naive Bayes (NB) [28], and Multi-layer Perceptron (MLP) [25]. Specifically, The Boolean function $L(E_s, E_t)$ was used to determine the existence of edge(s) between nodes E_s and E_t, where $L(E_s, E_t) = 1$ indicates positive links and $L(E_s, E_t) = 0$ represents negative links. We fit features of edge embeddings with labels provided by $L(E_s, E_t)$ to train the model. For positive examples, for each of the four networks, 60%, 10%, and 30% of all their edges were used for training, validation, and testing purposes respectively. For negative examples, an equal number of node pairs were randomly sampled (with the same ratio among training, validation, and testing sets as 60%, 10%, and 30% respectively).

For each classifier, we plotted the receiver operating characteristic (ROC) curve and computed the area under the ROC curve (ROAUC) to report link prediction performance. Moreover, as shown in Equations 9-12, we used precision, recall, F1 score, and average precision to quantify the link prediction performance amongst four edge embeddings.

$$Precision = \frac{|\{True\ Relations\} \cap \{Predicted\ Relations\}|}{|\{Predicted\ Relations\}|} \qquad (9)$$

$$Recall = \frac{|\{True\ Relations\} \cap \{Predicted\ Relations\}|}{|\{True\ Relations\}|} \quad (10)$$

$$F1\ score = \frac{2 * precision * recall}{precision + recall} \quad (11)$$

$$AP = \sum_n (Recall_n - Recall_{n-1}) Precision_n \quad (12)$$

For qualitative evaluation, we first visualized the network embeddings clustering output and used the silhouette score to evaluate clustering outputs. Silhouette score is adopted to calculate the average distance to entities in the same cluster with the average distance to entities in other clusters. Given any entity node e in cluster C_e, the internal mean distance is defined as:

$$m(e) = \frac{1}{|C_e| - 1} \sum_{t \in C_e, e \neq t} d(e,t) \quad (13)$$

where $d(e,t)$ is the distance between node e and t in C_e. Similarly, external mean distance is described as:

$$n(e) = \min_{k \neq e} \frac{1}{C_k} \sum_{t \in C_e} d(e,t) \quad (14)$$

Overall, the silhouette score is calculated incorporating both internal and external mean distances:

$$s(e) = \begin{cases} 1 - \frac{m(e)}{n(e)}, if\ m(e) < n(e) \\ 0, \quad if\ m(e) = n(e) \\ \frac{n(e)}{m(e)} - 1, if\ m(e) > n(e) \end{cases} \quad (15)$$

For some selected coronavirus infectious diseases, we also located the cluster they belonged to and checked the most similar entities within the same cluster using cosine similarity. Let E_s denote any given coronavirus infectious disease and E_t denote any biomedical entity inferred by network embeddings, and $f(E_s)$ and $f(E_t)$ represent the embeddings for E_s and E_t respectively, cosine similarity was calculated as shown in Equation 16.

$$cosine\_similarity\ (E_s, E_t) = \frac{f(E_s) \cdot f(E_t)}{\| f(E_s) \| \| f(E_t) \|} \quad (16)$$

Results

Different embeddings were generated by using neighbor size of 10, number of walks of 10, window size of 10, and dimensionality of 128. The optimal p and q were also tuned as 0.5 and 0.25 for each training process respectively. Detailed network embeddings and the corresponding clustering information could be found at: https://github.com/shenfc/COVID-19-network-embeddings. We have also implemented a web-based user-friendly tool for clustering visualization and entity similarity checking (https://www.davidoniani.com/covid-19-network).

Quantitative evaluation

As shown in Table 1, we presented the evaluation results of four different edge embedding operations along with six different classification algorithms. We found that, in general, the embeddings trained by the Average operation achieved the best performance across all the evaluation metrics. The optimal average precision, ROC score, precision and recall was reached when the RF was used, and the optimal F1 score was achieved by using NB. L1 and L2 had roughly the same performance, both peaking at ROC = 0.95 and AP = 0.96 with RF classifier. Among all four approaches, Hadamard yielded the worst performance, peaking at ROC = 0.89 and AP = 0.92 with RF classifier. Across all six classification algorithms, the worst performance, on the other hand, was shown by DT and MLP classifiers.

Regarding different classification algorithms, for the L1 embeddings embedding operation, RF and LR had similar performance, with ROC_RF = 0.96, AP_RF = 0.97 and ROC_LR = 0.96, AP_LR = 0.95. SVM, NB, and MLP were also not much different from each other. DT had the worst performance with ROC = 0.80 and AP = 0.75. Similarly, for L2, RF and LR had roughly the same performance. SVM, NB, and MLP were also equally performant. DT had the worst performance with ROC = 0.80 and AP = 0.75. For the Average operation, RF has outperformed all the other classification methods with ROC = 0.96 and AP = 0.97. LR and NB had similar performance. DT, SVM, and MLP were similar, yet all of them were behind RF, LR, and NB. For Hadamard, RF has shown the best performance with ROC = 0.89 and AP = 0.92. The rest of the methods have shown roughly the same performance except for

MLP which has the worst performance across all classification algorithms as well as across every embedding generation approach, with ROC = 0.60 and 0.56.

We finalized the network embeddings generated by the Average operation as the optimal one and a ROC curve for the performance across six classification algorithms is shown in Figure 2.

Table 1. Evaluation results for the four edge embeddings operations along with six machine learning algorithms (the highest value is marked in bold).

Operations	Algorithms	Average	ROC score	Precision	Recall	F1 score
Hadamard	DT	0.79	0.82	0.84	0.82	0.81
	LR	0.89	0.83	0.86	0.82	0.81
	SVM	0.80	0.81	0.85	0.81	0.81
	RF	0.92	0.89	0.87	0.86	0.86
	NB	0.82	0.84	0.86	0.84	0.84
	MLP	0.56	0.60	0.63	0.60	0.57
Average	DT	0.81	0.84	0.85	0.84	0.84
	LR	0.94	0.92	0.87	0.85	0.85
	SVM	0.83	0.86	0.87	0.86	0.85
	RF	**0.97**	**0.96**	**0.91**	**0.91**	0.90
	NB	0.88	0.91	**0.91**	**0.91**	**0.91**
	MLP	0.78	0.84	0.84	0.84	0.84
L1	DT	0.75	0.80	0.80	0.80	0.80
	LR	0.95	0.94	0.89	0.89	0.89
	SVM	0.87	0.89	0.90	0.89	0.89
	RF	0.96	0.95	0.89	0.88	0.88
	NB	0.85	0.88	0.89	0.88	0.88
	MLP	0.87	0.89	0.89	0.89	0.89
L2	DT	0.75	0.80	0.80	0.80	0.80
	LR	0.94	0.93	0.89	0.88	0.88
	SVM	0.87	0.88	0.90	0.88	0.88
	RF	0.96	0.95	0.89	0.88	0.88
	NB	0.85	0.87	0.88	0.87	0.87
	MLP	0.85	0.87	0.88	0.87	0.87

Figure 2. ROC scores for the Average operation with six machine learning algorithms.

Qualitative evaluation

We clustered the network embeddings by selecting the optimal hyperparameters $\varepsilon = 1.5$ and $k = 37$ for the DBSCAN algorithm. 63 clusters were generated with a Silhouette score of 0.128.

We used the network embeddings generated through the optimal Average operation to conduct a further qualitative evaluation. We first visualized clusters for diseases and clusters for all the entities as shown in Figure 3. In Figure 3(A), we found that *pneumonia*, *fever*, *fibrosis*, and *bronchitis* appeared in each cluster, indicating that they are common comorbidities amongst all types of coronavirus infectious diseases. We also observed that *COVID-19* co-

occurred with *coronavirus, SARS*, *Ebola*, and *ZIKA* in different clusters respectively, denoting there exist tremendous overlap between these diseases regarding underlying mechanisms. Clusters shown in Figure 3(B) further illustrate how different genes, mutations, and chemicals can link diseases with similar mechanisms. For example, *COVID-19* was grouped in cluster #6, which contains 606 biomedical entities in total, including *infection of SARS*, *ebola viruses*, *rs180047* mutation, and *carbohydrates* chemical component. Based on literature search, we found that *rs180047* is strongly related to *TGF-β1*, a master regulator for *pulmonary fibrosis*, which is a common comorbidity related to *COVID-19*, *infection of SARS*, and *Ebola viruses* [29]. In addition, *carbohydrates*-based diagnostic was recently reported to be a potential new approach for testing *COVID-19* [30], was also detected from cluster #6. The comprehensive information for entities included in each cluster was illustrated at: https://github.com/shenfc/COVID-19-network-embeddings.

Figure 3. Clustering visualization for diseases and all the biomedical entities. *COVID-19* is represented in red, *SARS* is represented in black, *coronavirus* is represented in green, *pneumonia* is represented in blue, *fever* is represented in cyan, *fibrosis* is represented in yellow, *diarrhea* is represented in magenta, *bronchitis* is represented in olivedrab, *Ebola* is represented in pink, *influenza* is represented in darkorchid, *ZIKA* is represented in khaki, all the genes are represented in purple, all the mutations are represented in silver, and all the chemicals are represented in salmon.

We then selected five coronavirus infectious diseases and listed top 10 closest entities using cosine similarity as shown in Table 2. *COVID-19* was clustered in cluster #6, and the top two closest entities in cluster #6 are *VP35* and *HD11* (both being genes). *VP35* is a virus protein of the other highly infectious disease *Ebola* [31]. *HD11*, also known as *Homeobox* protein, is known for regulating infectious diseases such as *Avian infectious bronchitis virus* (*IBV*) [32] that is in the coronavirus family. *Pulmonary coronavirus infection* was grouped in cluster #1 and has the closest associations with gene *PTP* (*protein tyrosine phosphatase*), which has been mentioned in *SARS-CoV* replication inhibition studies [33]. In case of *SARS-CoV-infected human airway epithelia cell cultures*, it is easy to notice that the entity is directly linked to the coronavirus infection [34]. As for *SARS-CoV infection damages lung* that listed in cluster #2, *IL-1-alpha* (*Interleukin 1Alpha*) also known as *IL-1α*, has the closest association with *SARS-CoV* and evidence could be found in a research study [35]. In particular, *IL-1α* is a *pro-inflammatory cytokine* that shows increase when infected by *SARS-CoV*. *Sucralfate* is a chemical compound that also holds close association with *SARS-CoV* , which has been studied as a potentially effective means against *Early-Onset Ventilator-Associated Pneumonia* [36]. *Coronavirus upper respiratory infection* was found in cluster #23, *pleuropneumoniae* (disease) and *plasmin* (gene) are most two similar entities. *Pleuropneumoniae* is a *pneumonia* complicated with *pleurisy*, which has been linked to *porcine upper respiratory tract* [37] and *Plasminogen* (*PLG*) (the *zymogen* of *plasmin*) has also been proved to be related with *Coronavirus upper respiratory infection* in research studies [38]. *Coronavirus-infected pneumonia* was detected in cluster #10. The top two closest diseases are *respiratory syncytial viral infection* and *pegylated interferon-alpha*, which could be proved by reviewing research studies in [39] and [40].

Table 2. Top 10 intra-cluster closest biomedical entities for five selected coronavirus infectious diseases. Cluster ID and the type of entities are marked in parentheses.

Coronavirus infectious diseases	Top 10 closest entities	Cosine similarity score
COVID-19 (Cluster #6)	VP35 (Gene)	0.9777
	HD11 (Gene)	0.9774
	Coronavirus infection process (Disease)	0.9700
	Fibroblast growth factor (FGF)-2 (Gene)	0.9655
	Acute respiratory infection illness	0.9596
	PIGS (Gene)	0.9576
	TGF alpha (Gene)	0.9571
	SFPQ (Gene)	0.9561
	Tumour necrosis factor (TNF) (Gene)	0.9549
	Praziquantel (Chemical)	0.9537
Pulmonary coronavirus infection (Cluster #1)	PTP (Gene)	0.9754
	SARS-CoV-infected human airway	0.9699

	"5'-tgg gat tca aca" (Chemical)	0.9672
	Tracheanasal respiratory epithelial cells	0.9658
	Suppressor of cytokine signaling 3	0.9620
	KAT (Gene)	0.9604
	CD32 (Gene)	0.9573
	Maternal SARS infection (Disease)	0.9553
	Respiratory syndrome coronavirus	0.9547
	S27 (Gene)	0.9546
Sars-cov infection damages lung (Cluster #2)	IL-1-alpha (Gene)	0.9560
	Sucralfate prn (Chemical)	0.9589
	Acute respiratory syndrome-cov	0.9555
	IL-5- and IL-13-producing ilc-iis (Gene)	0.9487
	HAP1 (Gene)	0.9342
	FSK (Chemical)	0.9337
	Low fever (Disease)	0.9328
	HIV and Ebola virus infection (Disease)	0.9327
	YKL-40 (Gene)	0.9288
	ETF (Gene)	0.9280
Coronavirus upper respiratory infection (Cluster #23)	Viruses actinobacillus	0.9890
	Plasmin (Gene)	0.9719
	JAM-1 (Gene)	0.9654
	TNF receptor-associated factor 6 (Gene)	0.9648
	GPC3 (Gene)	0.9613
	Renin (Gene)	0.9582
	ZO-1 (Gene)	0.9563
	Cathepsin G (Gene)	0.9556
	rs5743313 (Mutation)	0.9547
	Alpha1 antitrypsin (Gene)	0.9544
Coronavirus-infected pneumonia (Cluster #10)	Respiratory syncytial viral infection	0.9923
	Pegylated interferon-alpha (Chemical)	0.9891
	IFITM6 (Gene)	0.9872
	Feline b (Chemical)	0.9858
	E119V (Mutation)	0.9854
	Epac2 (Gene)	0.9850
	GFTP2 (Gene)	0.9849
	Hepatitis coronavirus infection (Disease)	0.9843
	Ouabain (Chemical)	0.9797
	LY6G (Gene)	0.9786

Discussion and future work

In this study, we used 11 keywords to query CORD-19-on-FHIR for constructing the COVID-19-centered coronavirus co-occurrence network. As research studies of COVID-19 published on a daily basis, we will keep watching the new results and adding more significant diseases/comorbidities into the keyword list to provide timely update for the co-occurrence network.

PMID information was not incorporated as an attribute into the co-occurrence network, which creates difficulties on the capability to trace each biomedical entity back to the original literature. In the future, we will add PMID list for each entity. On one hand, it will help our evaluation on detecting if the closely associated terms are from the same article or different publications. On the other, it will provide more evidence to scientists and clinical investigators for assisting their research studies on COVID-19 in an efficient manner.

We didn't consider the co-occurred frequency for the pairs of biomedical entities while training the network embeddings. Instead of treating each edge equally, in the future, we will add weights over edges using frequency in order to better represent associations in the network and provide more accurate edge embeddings to better quantify power of associations amongst biomedical entities.

In unsupervised learning approach, there is always a balance between the number of clusters and the silhouette score. In most cases, silhouette score tends to be higher if the number of cluster is small [41]. In this study, we used a heuristic way to determine silhouette score and the number of clusters for making clear separations over the biomedical entities. In the future, we sought to use our previously developed hierarchical clustering optimization algorithms to make dynamic balance between the optimal silhouette score and suitable cluster density [42-44]. Moreover, after checking top similar entities for five selected coronavirus infectious diseases, we observed that applying clustering over the network embeddings could detect both explicit and implicit associations. It is easy to check explicit links as most of them might be documented in existing studies. But for implicit associations, although they might hold huge potential on new discoveries, in order to validate their correctness, we will invite a clinical investigator from Mayo Clinic Division of Pulmonary and Critical Care Medicine for manual evaluation.

Conclusion

This study has explored the construction of co-occurrence network embeddings for COVID-19 and related coronavirus infectious diseases. We have tested different edge embeddings operations along with different machine learning algorithms to optimize the final network embeddings and developed unsupervised clustering algorithms to deep dive into specific COVID-19 related associations. Results indicated that the co-occurrence network embeddings were able to perform link prediction task well and detect both explicit and implicit associations for COVID-19, demonstrating its potential usage for discovering new disease management and treatment plan for COVID-19. Detailed implementations and data sources could be found at: https://github.com/shenfc/COVID-19-network-embeddings. A web-based user-friendly tool for clustering visualization and entity similarity check is available at: https://www.davidoniani.com/covid-19-network.

Funding Statement

This work was supported by the National Institute of Health (NIH) grants U01TR0062-1.

Author Contributions

DO was responsible for algorithm development, tool implementation, and manuscript drafting. GJ was responsible for extracting needed information from CORD-19-on-FHIR. HL provided supports on system evaluations. FS conceived and supervised the study. FS was responsible for research topic formulation, algorithm development, experiment design, tool implementation, and manuscript writing & revision.

Competing Interests Statement

None declared.

Acknowledgement

We thank the FHIRCat team for building the CORD-19-on-FHIR datasets and provide support on query template design.

Reference:

[1] Chen Q, et al. Keep up with the latest coronavirus research. Natur. 2020;579:193-.
[2] Mihalcea R, et al. Corpus-based and knowledge-based measures of text semantic similarity. Aaai2006. p. 775-80.
[3] Oliva J, et al. SyMSS: A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering. 2011;70:390-405.
[4] Heath T, et al. Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology. 2011;1:1-136.
[5] Ahamed S, et al. Information Mining for COVID-19 Research From a Large Volume of Scientific Literature. arXiv preprint arXiv:200402085. 2020.
[6] Bellomarini L, et al. COVID-19 and Company Knowledge Graphs: Assessing Golden Powers and Economic Impact of Selective Lockdown via AI Reasoning. arXiv preprint arXiv:200410119. 2020.
[7] Tsiotas D, et al. The effect of anti-COVID-19 policies to the evolution of the disease: A complex network analysis to the successful case of Greece. arXiv preprint arXiv:200406536. 2020.
[8] CORD-19. Available at: https://allenai.org/data/cord-19 (Accessed in May 2020).
[9] Wolinski F. Visualization of Diseases at Risk in the COVID-19 Literature. arXiv preprint arXiv:200500848. 2020.
[10] Wang X, et al. Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision. arXiv preprint arXiv:200312218. 2020.
[11] CORD-19-on-FHIR. Available at: https://github.com/fhircat/CORD-19-on-FHIR. (Accessed in May 2020).
[12] Bender D, et al. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE international symposium on computer-based medical systems: IEEE; 2013. p. 326-31.
[13] Miller E. An introduction to the resource description framework. Bulletin of the American Society for Information Science and Technology. 1998;25:15-9.
[14] Groza A. Detecting fake news for the new coronavirus by reasoning on the Covid-19 ontology. arXiv preprint arXiv:200412330. 2020.
[15] Mikolov T, et al. Linguistic regularities in continuous space word representations. Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies2013. p. 746-51.
[16] Grover A, et al. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining2016. p. 855-64.
[17] Shen F, et al. Constructing Node Embeddings for Human Phenotype Ontology to Assist Phenotypic Similarity Measurement. 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W): IEEE; 2018. p. 29-33.
[18] Shen F, et al. HPO2Vec+: leveraging heterogeneous knowledge resources to enrich node embeddings for the human phenotype ontology. Journal of biomedical informatics. 2019;96:103246.
[19] Wei C-H, et al. PubTator: a web-based text mining tool for assisting biocuration. Nucleic acids research. 2013;41:W518-W22.
[20] Tang L, et al. Leveraging social media networks for classification. Data Mining and Knowledge Discovery. 2011;23:447-78.
[21] Fortunato S. Community detection in graphs. Physics reports. 2010;486:75-174.
[22] Maaten Lvd, et al. Visualizing data using t-SNE. Journal of machine learning research. 2008;9:2579-605.
[23] Sander J, et al. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery. 1998;2:169-94.

[24] Quinlan JR. Induction of decision trees. Machine learning. 1986;1:81-106.
[25] Walker SH, et al. Estimation of the probability of an event as a function of several independent variables. Biometrika. 1967;54:167-79.
[26] Cortes C, et al. Support-vector networks. Machine learning. 1995;20:273-97.
[27] Ho TK. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition: IEEE; 1995. p. 278-82.
[28] Rish I. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence2001. p. 41-6.
[29] Osadnik T, et al. The association of functional polymorphisms in genes encoding growth factors for endothelial cells and smooth muscle cells with the severity of coronary artery disease. BMC cardiovascular disorders. 2016;16:218.
[30] Carbohydrate-based Diagnostics: A New Approach to COVID-19 Testing?. Available at: https://www.technologynetworks.com/diagnostics/news/carbohydrate-based-diagnostics-a-new-approach-to-covid-19-testing-332313. (Accessed in May 2020).
[31] Fajardo-Ortiz D, et al. Hegemonic structure of basic, clinical and patented knowledge on Ebola research: a US army reductionist initiative. Journal of translational medicine. 2015;13:124.
[32] Li H, et al. Gga-miR-30d regulates infectious bronchitis virus infection by targeting USP47 in HD11 cells. Microbial pathogenesis. 2020;141:103998.
[33] Yap Y, et al. Structural analysis of inhibition mechanisms of aurintricarboxylic acid on SARS-CoV polymerase and other proteins. Computational biology and chemistry. 2005;29:212-9.
[34] Sheahan TP, et al. An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. Science translational medicine. 2020;12.
[35] Barnard DL, et al. Enhancement of the infectivity of SARS-CoV in BALB/c mice by IMP dehydrogenase inhibitors, including ribavirin. Antiviral research. 2006;71:53-63.
[36] Bornstain C, et al. Sedation, sucralfate, and antibiotic use are potential means for protection against early-onset ventilator-associated pneumonia. Clinical infectious diseases. 2004;38:1401-8.
[37] Sidibe M, et al. Detection of Actinobacillus pleuropneumoniae in the porcine upper respiratory tract as a complement to serological tests. Canadian Journal of Veterinary Research. 1993;57:204.
[38] Wang J, et al. Tissue plasminogen activator (tpa) treatment for COVID-19 associated acute respiratory distress syndrome (ARDS): a case series. Journal of thrombosis and haemostasis. 2020.
[39] Zou S, et al. FDG PET/CT of COVID-19. Radiology. 2020:200770.
[40] Haagmans BL, et al. Pegylated interferon-α protects type 1 pneumocytes against SARS coronavirus infection in macaques. Nature medicine. 2004;10:290-3.
[41] Shen F, et al. Knowledge discovery from biomedical ontologies in cross domains. PloS one. 2016;11.
[42] Shen F, et al. Biobroker: Knowledge discovery framework for heterogeneous biomedical ontologies and data. Journal of Intelligent Learning Systems and Applications. 2018;10:1-20.
[43] Shen F, et al. BmQGen: Biomedical query generator for knowledge discovery. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE; 2015. p. 1092-7.
[44] Shen F, et al. Predicate oriented pattern analysis for biomedical knowledge discovery. Intelligent information management. 2016;8:66.

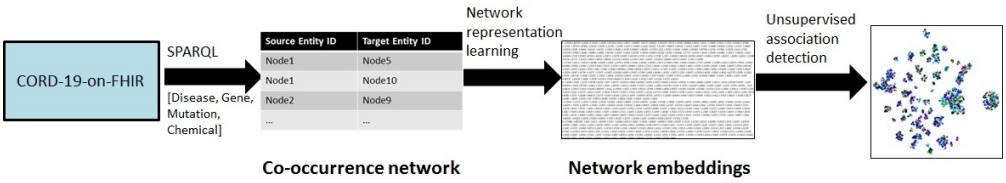

Figure 1. Study workflow.

311x60mm (96 x 96 DPI)

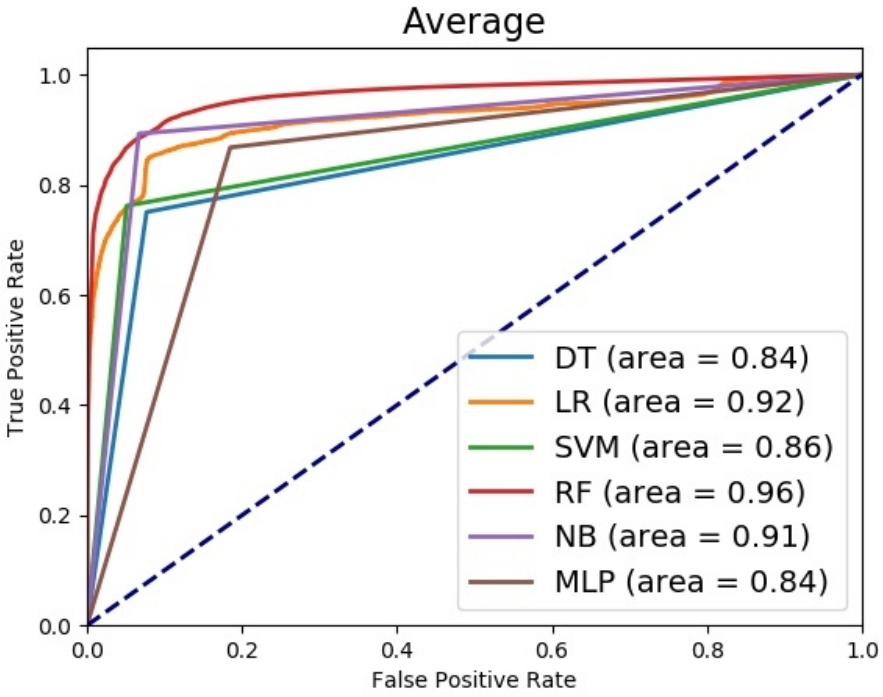

Figure 2. ROC scores for the Average operation with six machine learning algorithms.

169x127mm (96 x 96 DPI)

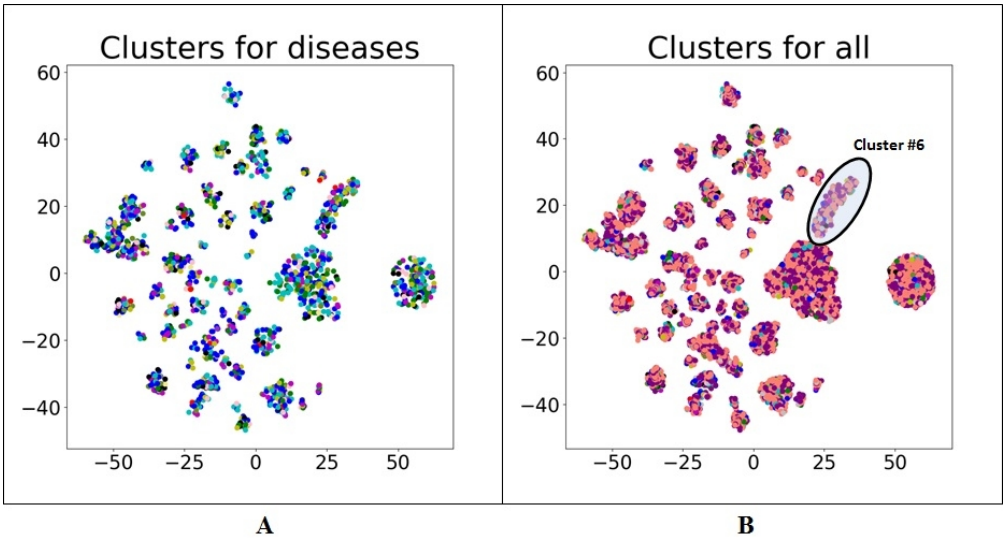

Figure 3. Clustering visualization for diseases and all the biomedical entities. COVID-19 is represented in red, SARS is represented in black, coronavirus is represented in green, pneumonia is represented in blue, fever is represented in cyan, fibrosis is represented in yellow, diarrhea is represented in magenta, bronchitis is represented in olivedrab, Ebola is represented in pink, influenza is represented in darkorchid, ZIKA is represented in khaki, all the genes are represented in purple, all the mutations are represented in silver, and all the chemicals are represented in salmon.

246x135mm (96 x 96 DPI)