Available online at www.sciencedirect.com

ScienceDirect

Biomedical Journal

journal homepage: www.elsevier.com/locate/bj

Original Article

Novel approach by natural language processing for COVID-19 knowledge discovery

Li Wang [a,d,1], *Lei Jiang* [b,1], *Dongyan Pan* [c,1], *Qinghua Wang* [a], *Zeyu Yin* [a], *Zijian Kang* [b], *Haoran Tian* [b], *Xuqiang Geng* [b], *Jinsong Shao* [e], *Wenjie Pan* [a], *Jian Yin* [b], *Li Fang* [p], *Yue Wang* [f], *Weide Zhang* [g], *Zhixiu Li* [h], *Jun Zheng* [i], *Wenxin Hu* [i], *Yunbao Pan* [j], *Dong Yu* [k], *Shicheng Guo* [l,m], *Wei Lu* [n], *Qiang Li* [o], *Yunyun Zhou* [p,2], *Huji Xu* [b,q,r,*,2]

[a] *Medical School, Nantong University, Nantong, China*
[b] *Department of Rheumatology and Immunology, Shanghai Changzheng Hospital, Second Military Medical University, Shanghai, China*
[c] *Department of Ophthalmology, Shanghai Changhai Hospital, Second Military Medical University, Shanghai, China*
[d] *Research Center for Intelligence Information Technology, Nantong University, Nantong, China*
[e] *Public Health School, Nantong University, Nantong, China*
[f] *Department of Histology & Embryology, Second Military Medical University, Shanghai, China*
[g] *Big Data and Artificial Intelligence Center, Zhongshan Hospital, Fudan University, Shanghai, China*
[h] *Translational Genomics Group, Institute of Health and Biomedical Innovation, Queensland University of Technology at Translational Research Institute, Princess Alexandra Hospital, Brisbane, Australia*
[i] *School of Data Science & Engineering, East China Normal University, Shanghai, China*
[j] *Department of Laboratory Medicine, Zhongnan Hospital of Wuhan University, Wuhan University, Wuhan, China*
[k] *Center for Translational Medicine, Second Military Medical University, Shanghai, China*
[l] *Department of Medical Genetics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA*
[m] *Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, USA*
[n] *NO.905 Hospital, Shanghai, China*
[o] *Department of Respiratory and Critical Care Medicine, Shanghai East Hospital, Tongji University, Shanghai, China*
[p] *Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, PA, USA*
[q] *Peking-Tsinghua Center for Life Sciences, Tsinghua University, Beijing, China*
[r] *School of Clinical Medicine, Tsinghua University, Beijing, China*

ARTICLE INFO

Article history:
Received 24 July 2020
Accepted 21 March 2022
Revised 25 February 2022

ABSTRACT

Background: The impact of COVID-19 on public health has mandated an 'all hands on deck' scientific response. The current clinical study and basic research on COVID-19 are mainly based on existing publications or our knowledge of coronavirus. However, efficiently

* *Corresponding author*. Department of Rheumatology and Immunology, Changzheng Hospital, Second Military Medical University, 800, Xiangyin Rd., Shanghai 200433, China.
E-mail address: xuhuji@smmu.edu.cn (H. Xu).

Peer review under responsibility of Chang Gung University.

[1] Li Wang, Lei Jiang and Dongyan Pan contributed equally to this work.
[2] Yunyun Zhou and Huji Xu contributed equally to this work and are co-corresponding authors.

https://doi.org/10.1016/j.bj.2022.03.011

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011

Keywords:
SARS-COV-2
ACE2
TMPRSS2
COVID-19
Natural language processing

retrieval of accurate, relevant knowledge on COVID-19 can pose significant challenges for researchers.

Methods: To improve quality in accessing important literature findings, we developed a novel natural language processing (NLP) method to automatically recognize the associations among potential targeted host organ systems, associated clinical manifestations, and pathways. We further validated these associations through clinician experts' evaluations and prioritize candidate drug targets through bioinformatics network analysis.

Results: We found that the angiotensin-converting enzyme 2 (ACE2), a receptor that SARS-CoV-2 required for cell entry, is associated with cardiovascular and endocrine organ system and diseases. Furthermore, we found SARS-CoV-2 is associated with some important pathways such as IL-6, TNF-alpha, and IL-1 beta-induced dyslipidemia, which are related to inflammation, lipogenesis, and oxidative stress mechanisms, suggesting potential drug candidates.

Conclusion: We prioritized the list of therapeutic targets involved in antiviral and immune modulating drugs for experimental validation, rendering it valuable during public health crises marked by stresses on clinical and research capacity. Our automatic intelligence pipeline also contributes to other novel and emerging disease management and treatments in the future.

At a glance of commentary

Scientific background on the subject

COVID-19 is an emerging disease causing the public health difficulties in recent years. A number of COVID-19 studies have been published in a short time, however there are limited automated tools to help researchers to collect relevant information from the ocean of literature.

What this study adds to the field

We developed a new NLP tool called LEI4COV to facilitate the knowledge discovery for COVID-19. Not limited for COVID-19, LEI4COV can be used to efficiently retrieval of relevant knowledges on the relationship of genes, pathways, and drugs for other emerging infectious disease through literature mining.

Early responses to the escalating emergence of human disease occasioned by infection with the novel coronavirus SARS-CoV-2 were hampered by confusion about the disease's symptoms and natural history, as well as its incidence [1]. At the initial stages of the COVID-19 pandemic, clinical manifestations were largely viewed as *mild*, with patients experiencing fever, coughing, vomiting, and headaches, or *severe*, with patients experiencing dyspnea, coagulation dysfunction, and potentially impaired function in the kidneys and other organ systems [2,3]. Because of the complexity of clinical manifestations, precise diagnosis and treatment for COVID-19 have been an ongoing challenge. Additionally, since there are no effective drugs to treat COVID-19, finding drugs based on coronavirus pathogenesis, clinical manifestations, organ involvement and past treatment experience is urgent.

A number of studies used bioinformatics approaches to predict COVID-19 candidate drugs from gene expression analyses and protein–protein interaction analyses [4,5]. Other studies have focused on predicting drug candidates through virtual docking screening of molecular 3D structures [6–8]. These approaches have primarily used systems biology approaches to identify COVID-19 drug candidates. Artificial intelligence (AI)-based natural language processing (NLP) approaches provide a new avenue for pinpointing evidence-based medicine potentially able to thwart coronavirus pathogenesis from a large number of unstructured research articles. NLP can automatically elucidate hidden knowledge in the textual representations of biomedical concepts from the literature. However, to date, no published studies have comprehensively studied COVID-19 clinical manifestations and biomedical mechanisms to prioritize therapeutic targets using NLP approaches.

To achieve this goal, we developed an approach to identify literature evidence information for COVID-19 (LEI4COV). Our iteration of the LEI4COV method is based on a deep learning framework with an advanced self-attention structure. This state-of-the-art algorithm has proven its strength to better detect contextual relations of entities [9]. LEI4COV is able to learn correlated biomedical concept entities from a large amount of unstructured text from the literature by converting entities into high-dimensional vector space by maximizing their co-occurrence probabilities, and in this way capturing closely related entities within a smaller distance in vector space. LEI4COV enables biology researchers and clinicians to efficiently test their hypothesis or to run in silico experiments for the discovery of important knowledge from the literature. Not limited to COVID-19, LEI4COV can be used to infer the biomedical knowledges for other diseases.

It is known that coronaviruses enter cells through the binding of their viral spike (S) proteins to the host's cellular receptors [10]. The S protein is actually primed by host cell proteases, so blocking the host's receptor proteins and their

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011

helpers could stop the entry of the virus into cells [11]. Recent studies have confirmed that SARS-CoV-2 makes use of the *SARS-CoV receptor ACE2* to enter cells and the *serine protease TMPRSS2* to activate the binding of viral S proteins [12–14]. Given this recently acquired knowledge, we reasoned that using LEI4COV to identify evidence-based reports on *ACE2 and TMPRSS2* might rapidly and efficiently yield important knowledge on the relationship of these genes to SARS-CoV-2. This knowledge could help inform efforts to effectively target receptor proteins and their helpers to control infections in COVID-19.

In this project, we used LEI4COV to access and categorize a broad compendium of current knowledge and insights on COVID-19, particularly those associated with (1) clinical manifestations, (2) target organs, (3) signaling pathways, and (4) predictions on the repurposing drugs to treat COVID-19. Our findings provide valuable information on the use of data mining approaches such as ours to inform the intersection of precision medicine with public health. It also offers references of use as clinical and research communities are asked to rapidly – almost instantaneously – develop effective, viable, and affordable approaches for managing novel coronavirus-related diseases.

Material and methods

Project design and data collection

We used COVID-19 target genes *ACE2* and *TMPRSS2* as keywords to conduct literature retrieval in the PubMed public database. Next, we used the NLP embedding approach to analyze ACE2- and TMPRSS2-related biomedical concepts, including organ systems, disease, and genes. We also performed a meta-analysis of coronavirus-infected cases and compared the ACE2- and TMPRSS2-related clinical manifestations. Then, we performed pathway enrichment analysis to identify COVID-19 gene-related pathways. Finally, we used a network analysis method, random walk with restart algorithm to prioritize drugs.

Natural language processing to extract clinical and biomedical concepts from unstructured free text

We used MetaMap to extract biomedical concept terms from unstructured literature. MetaMap was published by the National Library of Medicine (NLM) in 2001 and is considered the foundation for biomedical NLP tools for information extraction by mapping biomedical text to the Unified Medical Language System (UMLS) Metathesaurus [15]. MetaMap uses computational-linguistic techniques and has been widely used for semiautomatic and fully automatic indexing of biomedical literature at NLM. Since the initial information extracted by MetaMap contained too many redundant terms not related to research purposes, we only kept terms belonging to organ systems, diseases, and genes. Since some genes had aliases, we normalized gene names using GeneCards [16]. All medical entities were tokenized as the input of the vector for downstream embedding analysis.

NLP embedding approach to identify clinical manifestations associated with COVID-19 genes

We developed a new NLP embedding approach to identify COVID-19-related biomedical concepts from the literature evidence information (LEI4COV: https://github.com/hitales-tech/Lei4Cov). This approach is an extended version of our recently published work EHR2Vec, which was designed for vector embedding on electronic health record (EHR) clinical notes [17]. EHR2Vec integrated the word embedding algorithm Word2Vec with a multi-head self-attention structure, which has shown its improved performance of the embedding accuracy compared to other representation learning approaches [9]. LEI4COV applied the same DL algorithm as EHR2Vec but with a particular emphasis on the representations learning for biomedical concepts from literature. Instead of analyzing the clinical note per visit in EHR2Vec, LEI4COV used each abstract as the analysis unit and performed the self-attention mechanism analysis within each abstract window.

The initialized vector-matrix W is in vector space R^{h*c}, where c is the dimension of each entity vector and h is the number of entities in all abstract. Here, we used default value $c = 512$, which means each entity maps to 512-dimensional vector space. This hyper-parameter is chosen based on the trade-off of computational complexity and accuracy based on our previous experiments. We first input the initialized vector matrix to the first sublayer (attention mechanism). Then, the $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}((QK^T)/\sqrt{d_k})V$ is the core formula of the attention mechanism that is used, in which Q, K and V represent the query vector, key vector, and value vector, respectively, and d_k represents the dimension of Q, K or V [9]. In the multi-head attention model, a total of eight attention heads were used as the default. The eight attention heads are equivalent to eight subtasks, each subtask generating its own attention with in each abstract window.

The optimized vector matrix W is obtained through iterative training. We obtained the final matrix by continuously optimizing the vector-matrix W. Assume e_i, e_j represent the different entities in the abstract, E_t represents one abstract. The co-occurrence log-likelihood function is used to optimize the obtained vectors (Eq. (1)).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i:e_i\in E_t}\sum_{j:e_j\in E_t, j\neq i}\log p(e_j|e_i),\ \text{where},\ p(e_j|e_i)=\frac{\exp\left(W[i,:]^T W[j,:]\right)}{\sum_{k=1}^{all}\exp\left(W[k,:]^T W[i,:]\right)} \tag{1}$$

Since some of the abstracts have extremely long entity sequences, we used the 98% quantile of max length as the

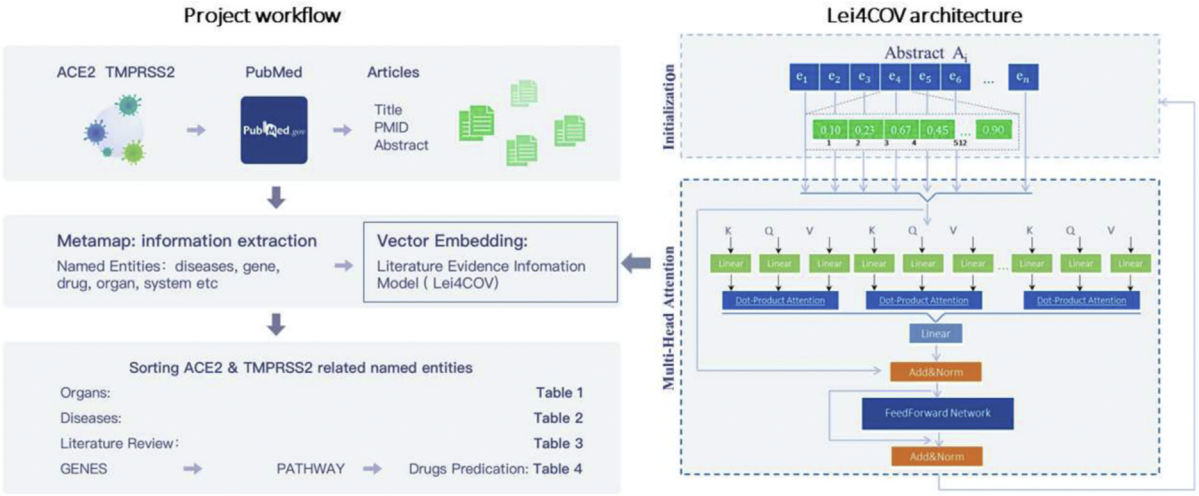

Fig. 1 Project design overview. We searched for studies related to two COVID-19 genes, ACE2 and TMPRSS2, in PubMed. Next, we used NLP methods to extract biomedical concepts and calculated their correlations with COVID-19 genes. Then, we filtered out unrelated entities and only kept concept entities belonging to organs, disease, genes, etc. Finally, we performed pathway enrichment analysis and predicted drugs by network analysis.

cutoff as the entity sequence length. The entity sequences from abstracts that are longer than the 98% quantile of the maximum value were truncated and shorter were padded with 0. Finally, associations between entity vectors were further calculated by the cosine similarity score (Sc).

LEI4COV was implemented and trained using the TensorFlow 1.8.0 deep learning framework. All models were performed on a CentOS server equipped with two 16G NVIDIA TESLA P100 graphics cards. LEI4COV used the Adadelta optimizer to optimize the target function with a drop rate of 0.1 to achieve model convergence. LEI4COV used eight attention heads in the self-attention mechanism, 512 vector dimensions for each entity, and trained for 20 epochs to obtain the best result.

Manual synthesis of literature evidence by clinician experts to confirm the clinical manifestations for COVID-19

To test the reliability of our analysis, we performed a literature review of coronavirus-infected cases. Keywords including SARS-CoV-2, 2019-nCoV, COVID-19, SARS-CoV, SARS, MERS-CoV, and MERS were used to conduct initial literature retrieval in PubMed and CNKI. The literature included 1278 cases of COVID-19 (3 references), 7671 cases of SARS (12 references), and 608 cases of MERS (8 references). Two experienced physicians independently reviewed the related literature concerning clinical presentation/clinical characteristics/clinical features/symptoms/clinical manifestation. The number of symptoms was counted, and the symptoms of each system were counted and calculated as percentages.

Network analysis to prioritize repurposing drug candidates for COVID-19

Since there was no standard threshold for the cosine similarity score (Sc) between queried vectors (*ACE2* and *TMPRSS2*) and other molecule vectors, we set the top 200 as the cutoff to consider them as the most relevant NERs. These correlated genes were subjected to pathway enrichment analysis using the direct interaction algorithm of MetaCore™ (Clarivate Analytics) [18]. According to the results, pathways were verified in the 21 reported drugs of 2019-nCoV based on host-based treatment strategies [19]. In addition, the top 50 pathway-related MetaCore-collected drugs were summarized to support further clinical trials. We used the PageRank algorithm to prioritize drugs for both *ACE2*- and *TMPRSS2*-related genes [20]. PageRank, which was developed by Google, is a network-based approach used to rank the most important web pages. Here, it works by counting the number and quality of pathways linked to a drug in a network to determine how important the drug is. The underlying assumption is that more important drugs are likely to receive more links from pathways. The PageRank score and degree of network connectivity were calculated by the R igraph package [21].

Results

LEI4COV enabled rapid acquisition of representative entities associated with COVID19

An overview of our project design can be found in Fig. 1. We collected a total of 1912 abstracts related to *ACE2* published between 1994 and 2020 and a total of 1025 abstracts related to *TMPRSS2* published between 1997 and 2020. Using the NLP information extraction tool MetaMap, a total of 15,845 biomedical concept entities were extracted from unstructured literature with diverse semantic categories, such as diseases and genes [15]. These entities were converted to digital vectors by our vector embedding approach, LEI4COV, with cosine similarity scores between entities represented as Sc. Our assumption is if LEI4COV converted digital vectors correctly, the vector embeddings should be able to separate abstracts into two classes, ACE2-related or TMPRSS2-realated abstracts.

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011

Table 1 Top 10 correlated organs/systems for ACE2 and TMPRSS2.

ACE2		TMPRSS2	
Organ/system entities	Sc	Organ/system entities	Sc
Renin–angiotensin system	0.85	Prostate	0.83
Heart	0.68	Hippocampus proper	0.39
Kidney	0.68	Lung	0.21
Axis vertebra	0.55	Urinary tract	0.20
Cardiovascular system	0.51	Bony process	0.20
Renin–angiotensin–aldosterone system	0.47	Lutheran blood-group system	0.19
Lung	0.42	Hand	0.19
Brain	0.41	Exocrine glands	0.18
Bony process	0.37	Gland	0.17
Blood vessel	0.33	Region of prostate	0.16

We comparatively studied the embedding performance of LEI4COV with other state-of-art algorithms, Word2Vec and BioBERT. Results in Supplementary Fig. 1 showed LEI4COV performs best among the three embedding algorithms. Therefore, we provided the initial top 200 most relevant medical concept entities for *ACE2* and *TMPRSS2* in Supplementary Tables S1–2. After further excluding some irrelevant entities, we only kept entities from three categories of interest, including molecular/protein entities, organ/system entities, and disease entities. For ACE2, we retrieved a total of 81 molecular/protein entities, 32 organ/system entities, and 38 disease entities. For TMPRSS2, *LEI4COV* retrieved a total of 86 molecular/protein entities, 21 organ/system entities, and 19 disease entities.

LEI4COV enabled identification of clinical manifestations relevant to COVID-19

LEI4COV allowed us to quickly identify that *ACE2* and *TMPRSS2* are associated with different organ systems [See Table 1]. *ACE2* appears more relevant within cardiovascular organ systems, such as the renin–angiotensin system, heart, and kidneys. *TMPRSS2*, by contrast, appears more relevant to the prostate, hippocampus proper, and lungs. For general disease associations, we found that both *ACE2* and *TMPRSS2* were highly associated with communicable diseases [Supplementary Table S3]. When we compared virus-related disease associations, we found that *ACE2* and *TMPRSS2* were highly associated with SARS coronavirus, as well as other coronaviruses, as shown in Supplementary Table S4. These results were consistent with clinical observations.

An overview and summary comparison of clinical manifestations associated with ACE2 and TMPRSS2 are provided in Fig. 2. ACE2, in addition to its association with viral infection, is distinctly associated with cardiovascular disease, in which hypertensive disease, at 0.67, has the highest correlation with ACE2, with at, cardiac arrest next at 0.56, and heart failure at 0.43. Endocrine system disease was also common in ACE2-associated disease, while infectious disease was common in TMPRSS2-related disease. One explanation could be that ACE2 was also a key enzyme in the renin–angiotensin–aldosterone system and was a target for cardiovascular disease treatment. In addition, ACE2 could increase blood flow perfusion of islets, thus improving secretion of islets and contributing to diabetes control. This might explain why cardiovascular disease and endocrine system disease were common in ACE2-associated disease [Fig. 3].

COVID-19 clinical manifestations confirmed by human experts' meta-analysis

To further confirm the relevance of clinical manifestations with COVID-19 found by LEI4COV, a total of 23 reference papers that, taken together, examined 9557 cases of coronavirus infection were retrieved from the PubMed database. The reports represented cases of SARS, MERS, and COVID-19 infections. Two study clinician experts (Tian, Geng) analyzed all symptoms described for the reported coronavirus cases. As seen in the statistics provided in Table 2, the clinical

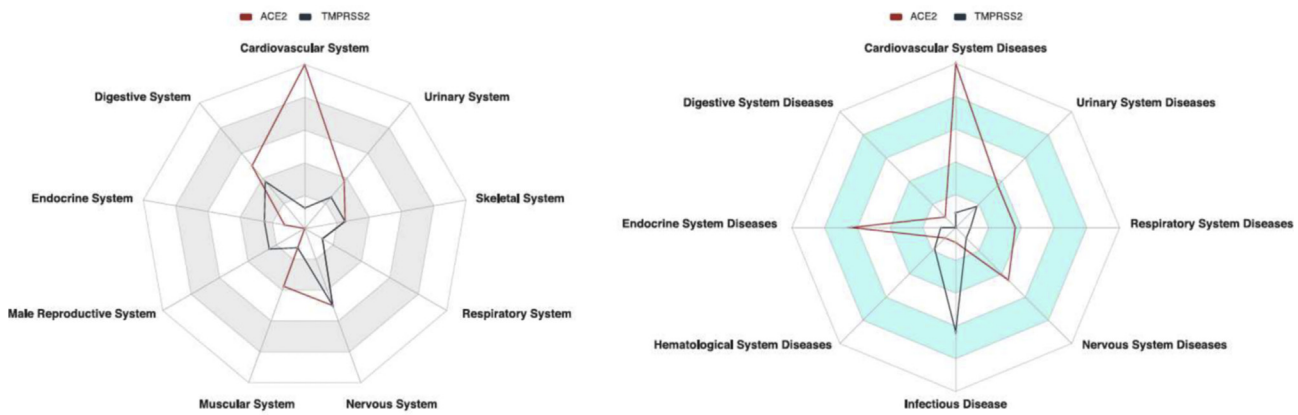

Fig. 2 Comparison of clinical manifestations for ACE2 and TMPRSS2. Radar plots showing the characteristics that were significantly different in ACE2- and TMPRSS2-related organs/systems and diseases. For ACE2-related organs/systems, the outermost point with the greatest value is the cardiovascular system. For ACE2-related disease, the outermost points are cardiovascular system disease and endocrine system disease. For TMPRSS2-related disease, the outermost point is infectious disease.

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011

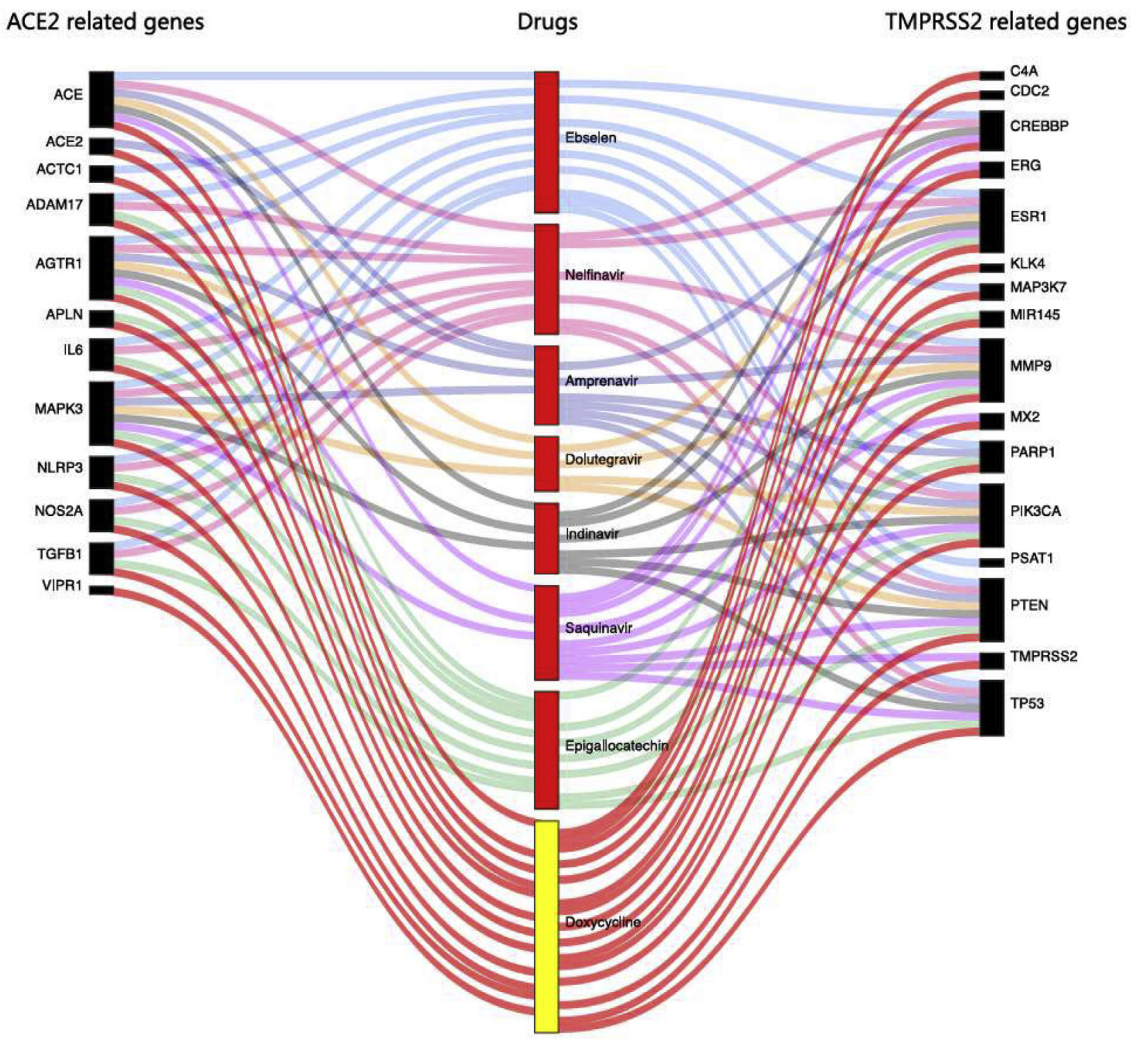

Fig. 3 Illustration of COVID-19 genes as the therapeutic targets for drugs. ACE2- and TMPRSS2-related molecular/protein entities were normalized to official gene symbols by GeneCards. After performing gene set enrichment analysis, some genes enriched in certain COVID-19 pathogenic pathways were identified as the therapeutic targets that link to potential drug candidates.

manifestations in these reports yielded three dominant symptom areas: respiratory system, digestive system and fever.

For COVID-19, SARS, and MERS, respiratory system involvement accounted for 42.9%, 30.7% and 50.4% of virus-related symptoms, respectively. That is, the impact of MERS on respiratory symptoms was higher than that of the other two coronavirus diseases. Digestive system involvement accounted for 3.5%, 10.1% and 14.2% of symptoms, respectively. Fever symptoms accounted for 28.0%, 29.1% and 20.8%, respectively. Our analysis by LEI4COV showed a similar involvement of respiratory and digestive system impact, with

Table 2 Clinical symptoms of coronavirus infection.

	COVID-19	SARS	MERS	Total
Cases	1278	7671	608	9557
Symptoms	4080	22,293	2145	28,518
Respiratory system	1751 (42.9%)	6842 (30.7%)	1082 (50.4%)	9675 (33.9%)
Digestive system	144 (3.5%)	2254 (10.1%)	305 (14.2%)	2703 (9.5%)
Fever	1141 (28.0%)	6479 (29.1%)	447 (20.8%)	8067 (28.3%)
Headache	153 (3.8%)	1944 (8.7%)	63 (2.9%)	2160 (7.6%)
Myalgia	211 (5.2%)	2039 (9.1%)	119 (5.5%)	2369 (8.3%)
Fatigue	533 (13.1%)	179 (0.8%)	127 (5.9%)	839 (2.9%)
Shiver	125 (3.1%)	2218 (9.9%)	1 (0)	2344 (8.2%)
Vertigo	13 (0.3%)	338 (1.5%)	0 (0)	351 (1.2%)
Conjunctivitis	9 (0.2%)	0 (0)	1 (0)	10 (0)

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011

Table 3 Top 3 predicted drugs in each category.

Candidate drugs	Sum score	ACE2 score	ACE2 PR score	TMPRSS2 score	TMPRSS2 PR score	Category
Doxycycline	74	44	0.030507557	30	0.031105069	Infection
Roxithromycin	53	41	0.027830773	12	0.013403104	Infection
Gemifloxacin	49	38	0.025035836	11	0.012095995	Infection
Doramapimod	37	24	0.017474146	13	0.014515815	Immune; Infection
Eritoran	11	10	0.007557487	1	0.002315731	Immune; Infection
Emricasan	11	2	0.002597711	9	0.010193351	Immune; Infection
Mycophenolic acid	53	39	0.025907587	14	0.015016035	Infection (antiviral)
Pimodivir	50	23	0.016885078	27	0.030772509	Infection (antiviral)
Oleic acid	49	29	0.020948295	20	0.021375538	Infection (antiviral)

PageRank (PR).

lung Sc = 0.4 and intestinal Sc = 0.21. Gastrointestinal symptoms were less common in patients with COVID-19 compared to SARS and MERS, which was also reported by others; the reason is still unknown [22].

COVID-19 clinical manifestations identified by LEI4COV are consistent with meta-analysis by human experts. By combining all cases from three diseases, we found that respiratory system involvement was the most common (45.7%), which was highly consistent with the respiratory system involvement predicted by our algorithm analysis with *ACE2* and TEMPRESS2 as the key words [Table 1 "lung" *ACE2* Sc = 0.42, TMPRSS2 Sc = 0.21]. Digestive system involvement was the second most prevalent symptom for all three diseases (13.4%), which was also consistent with the conclusion for ACE2 [see Table 1 "Intestines" Sc = 0.21]. Symptoms of the cardiovascular system (*ACE2* Sc = 0.33), kidney (*ACE2* Sc = 0.28), endocrine system (*ACE2* Sc = 0.25, TMPRSS2 Sc = 0.19) and brain (*ACE2* Sc = 0.23) were also found in the algorithm prediction, which showed that our algorithm had a reliable prediction performance. Additional results need to be confirmed by collecting more clinical data. Symptoms involving the cardiovascular system were seldom reported in the three coronavirus (CoV)-related clinical studies. One possible reason for this may be that cardiovascular symptoms were previously ignored, as some patients may have had coexisting cardiovascular disease before their coronavirus infection. Another reason might be that ACE2 is frequently a target for cardiovascular disease treatment, with cardiovascular disease therefore often presented in ACE2-related papers but not in reports of CoV-related diseases.

LEI4COV enabled identification of therapeutic targets in COVID-19 pathways

Meaningful molecules/proteins, including 31 ACE2-related and 45 TMPRSS2-related molecules/proteins, were screened. Examples of ACE2-associated molecules/proteins included VIPR1, ACE, SLC33A1, CTSC, APLN, ACHE, IL10, and IL-6. Official gene symbols for these entities, normalized by GeneCard [16], are provided in Supplementary Table S4. A gene set enrichment analysis of signaling pathways was performed by using a direct interaction algorithm of MetaCore (Clarivate Analytics) ($p < 0.05$). The TOP50 pathways of ACE2 and TMPRSS2 are listed in Supplementary Table S6.

These pathways, and their association with ACE2- and TMPRSS2-related molecules/proteins, might point toward effective drug candidates for COVID-19 treatment. For example, the pathways of TNF-alpha, IL-1 beta-induced dyslipidemia, and inflammation in obesity and type 2 diabetes in adipocytes ($p = 6.07E-10$), which are involved in inflammation, lipogenesis, lipolysis, fatty acid oxidation, and oxidative stress, suggest drug candidates such as mycophenolic acid and siltuximab. Additionally, the pathway of IL-6 signaling in colorectal cancer ($p = 1.79E-06$) suggests candidate drugs such as tocilizumab and nafamostat. Another interesting pathway that has gained attention is mucin expression. Several possibilities suggest exploration here. Proteases and EGFR-activated mucin production in airway epithelium in chronic obstructive pulmonary disorder (COPD) ($p = 3.17E-05$) may shed light on the considerable mucus accumulation in the airways of COVID-19 patients, which is widely observed in radiology studies and in autopsies [23].

Prioritized potential drugs for COVID-19 by network analysis

We searched all potential drugs related to the top 50 pathways for both ACE2 and TMPRSS2 in the MetaCore database. Removal of three overlapping pathways between the two genes yielded a combined list of 97 pathways linking to 427 potential drugs, 281 of which had Anatomical Therapeutic Chemical classification (ATC) codes. The 427 potential drugs were classified into five categories: infection, inflammation, immune, respiratory, and cancer. Among these drugs, several drugs are particularly interesting. Siltuximab, for example, classified as a cancer therapy, is a chimeric monoclonal antibody (mAb) that binds to and blocks the effect of IL-6. Recently, it was used to treat COVID-19 and reduced serum CRP levels to within the normal range by Day 5 [24]. Another cancer drug, tocilizumab, is an IL-6R-targeted mAb; it received rapid approval in China for the treatment of patients with severe COVID-19 with extensive lung damage [25]. Nafamostat, in the respiratory category, is a synthetic serine protease inhibitor similar to camostat, which was reported to partially block SARS-COV-2 S-driven entry into TMPRSS2+ cells and was considered a promising drug for COVID-19 treatment [14]. We also compared the 427 drug candidates predicted by our study with a recently published paper in Nature Drug Discovery, "Potential repurposing

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011

candidates for 2019-nCoV." The authors of this work proposed 21 host-targeted agents [19]. Analyzing the pathways of these 21 drugs, we found that 3 of these agents can be found in the ACE2 and TMPRSS2-related pathways, including oxidative stress_ROS-induced cellular signaling, the signal transduction_PTEN pathway, and transcription_androgen receptor nuclear signaling [Supplementary Table S7].

When we only focused on drug candidates belonging to the anti-infection category, our analysis was narrowed down to a total of 65 candidate drugs [Supplementary Table S8]. These drug candidates, ranked by their importance score using a Google PageRank algorithm, are shown in Table 3. Of them, doxycycline ranks as the most important drug for fighting. An antibiotic used to treat bacterial infections, doxycycline has also been found to have antiviral effects against vesicular stomatitis virus [26], dengue virus [27], and retrovirus [28], as well as attenuating acrolein-induced mucin production [29]. Doxycycline's effectiveness has not been indicated in COVID-19 patients, but it might be a promising candidate for COVID-19 treatment, given autopsy findings of considerable mucus in the alveoli of COVID-19 patients.

In our examination of agents useful for antiviral infection, mycophenolic acid topped the list. Mycophenolic acid, an immunosuppressant, is classified as a reversible inhibitor of inosine monophosphate dehydrogenase (IMPDH). In addition to its anti-inflammatory effects, it has also been reported to show an antiviral effect on coronaviruses [30] and has been proven effective against MERS in a small sample clinical report [31]. When we turned our focus to immune infection, eritoran, a synthetic TLR4 antagonist, ranked second. This drug can reportedly block influenza-induced lethality in mice, as well as decrease lung pathology, cytokine and oxidized phospholipid expression, and viral titers [32].

Discussion

Over the present century, a number of emerging viral pathogens have led to major public health disease outbreaks, such as SARS, MERS, H1N1, Ebola and COVID-19. Despite the century's advances in technology, these agents have greatly threatened social development, the economic fabric, and human life itself. The prevalence and severity of the present COVID-19 pandemic now highlight the urgent need for research on the pathogenic mechanisms, prevention strategies, diagnosis and treatment of the pathogens responsible for these outbreaks. However, scientific research grounded solely in traditional models of observation and hypothesis development may not be able to keep pace with the urgencies of epidemic prevention and control [1]. It is imperative instead to harness technological advances, not merely for computing power to establish statistical validity and findings but to assist in hypothesis generation. The use of technology to establish new, high-throughput approaches to the scientific research literature and the use of data mining tools based on natural language can significantly contribute to the urgent task of research and design, aimed at prevention and control, in the face of pandemic threats such as that now posed by a single, novel coronavirus.

In this study, the NLP approach was used to identify hidden relationships among biomedical concept entities through a mining of the literature. ACE2 and TMPRESS2, the key invasion target genes of COVID-19, were selected as the main key entries for related literature retrieval. Using a vector embedding approach, we ranked entities, including organ systems, disease, and the gene/proteins related to these two COVID-19 genes. Our findings thus support the insights provided by LEI4COV that can assist the rapid interrogation of pathways that could better inform clinical diagnosis, symptom detection and treatment across a rapidly escalating disease threat.

Not surprisingly, both *ACE2* and *TMPRESS* were associated with coronavirus-related disease. We also found that cardiovascular diseases, such as hypertension and diabetes, were particularly associated with ACE2. This finding was consistent with reports that COVID-19 has a high occurrence in patients with hypertension and diabetes [33]. Our findings showed LEI4COV's ability to help access references that can assist clinical diagnosis and understanding of a disease's natural history. Our study also included a literature review for meta-analysis by human experts. The findings of this expert review, which focused on symptom analysis for COVID-19, SARS, and MERS, verified the relevance of LEV4COV-identified studies, as the studies they included well-known respiratory and digestive system findings associated with coronavirus diseases.

We found several important molecules/proteins and pathways that were associated with *ACE2* and *TMPRSS2*. One was IL-6, which is involved in pathways such as the production and activation of TGF-beta in airway smooth muscle cells ($p = 1.42E-06$); IL-6 signaling in colorectal cancer ($p = 1.79E-06$); and the release of proinflammatory factors and proteases by alveolar macrophages in asthma ($p = 3.63E-06$). Another important finding was PTEN, which is involved in the signal transduction_PTEN pathway ($p = 2.27E-05$) and plays a role in proteases and EGFR-activated mucin production in airway epithelium in COPD ($p = 3.17E-05$). Examination of treatments used in diseases and disorders associated with these molecules, in turn, helped us identify candidate drugs for COVID-19 treatment consideration. Our approach also led to the identification of other potential drugs, successfully applied in treating infections having relevance to those occasioned by coronaviruses, such as mycophenolic acid, siltuximab, tocilizumab, nafamostat, and doxycycline. Baricitinib, an AP2-associated protein kinase 1 (AAK1) inhibitor that could interrupt virus passage into host cells and the intracellular assembly of virus particles, which we present in our drug candidate list, has also been predicted by others using BenevolentAI's knowledge graph tool, also developed from the NLP approach [34,35].

Limitations in the use of natural language document discovery are hampered by such factors as limitations on access to complete reports. Thus, most of our data mining was based on abstracts, with the data analyzed thus limited to core data discovery and incomplete mining a strong possibility. Additionally, the current data mining effort was predicated on two well-known COVID-19 genes, ACE2 and TMPRESS. As more knowledge of key functional genes and their roles in SARS-COV-2 are identified, more comprehensive and complete analyses will be possible. An important

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011

further consideration is that the mining strategy we employed was aimed at natural language and was thus greatly influenced by subjective conclusions drawn by study authors. In the future, it will be possible to weight the literature toward obtaining more refined sources by better screening tools. We are currently developing a more refined operation of tools and platforms that can improve the reliability and validity of findings accessed and included for analysis. Finally, as our results were based on NLP without experimental validation, further confirmation is needed by future studies.

Conclusion

In this project, we developed a novel NLP method LEI4COV to identify hidden relationships among biomedical concept entities through a mining of the literature. Using LEI4COV, we ranked the associated entities, including organ systems, disease, and the gene/proteins related to the key invasion target genes of COVID-19, ACE2 and TMPRESS2. Our results showed LEI4COV's ability to help access references that can assist clinical diagnosis. We found several important molecules/proteins and pathways that were associated with *ACE2* and *TMPRSS2*, such as IL-6 signal pathways and transduction_PTEN pathway. We also found several drug candidates such as siltuximab, tocilizumab, and doxycycline that need to be further validated. Our work can assist the rapid interrogation of pathways to inform clinical diagnosis and drug discovery.

Data and code availability

Data are available in the Supplementary Information. The code implementation is available at: https://github.com/hitales-tech/Lei4Cov.

Conflicts of interest

The authors declare no competing interests.

Acknowledgments

We thank Hitales (Shanghai, China) for algorithm and software engineering support and Clarivate Analytics (London, UK) for database support.

This work was supported by the Ministry of Science and Technology Key Research and Development Program of China (No. 2018YFC0116902) and the National Science Foundation of China (No. 81873915).

Appendix A. Supplementary data

Supplementary data to this article can be found online at 10.1016/j.bj.2022.03.011.

REFERENCES

[1] Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020;382:1708–20.
[2] Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA 2020;323:1061–9.
[3] Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020;395:507–13.
[4] Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 2020;6:14.
[5] Xing J, Shankar R, Drelich A, Paithankar S, Chekalin E, Dexheimer T, et al. Reversal of infected host gene expression identifies repurposed drug candidates for COVID-19. bioRxiv:2020.04.07.030734v1[Preprint]. 2020[cited ???]. Available from: https://www.biorxiv.org/content/10.1101/2020.04.07.030734v1
[6] Wang J. Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. Chem Inf Model 2020;60:3277–86.
[7] Stokes J, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia N, et al. A deep learning approach to antibiotic discovery. Cell 2020;180:688–702.
[8] Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B 2020;10:766–88.
[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008.
[10] Wang H, Yang P, Liu K, Guo F, Zhang Y, Zhang G, et al. SARS coronavirus entry into host cells through a novel clathrin- and caveolae-independent endocytic pathway. Cell Res 2008;18:290–301.
[11] Simmons G, Reeves JD, Rennekamp AJ, Amberg SM, Piefer AJ, Bates P. Characterization of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) spike glycoprotein-mediated viral entry. Proc Natl Acad Sci U S A 2004;101:4240–5.
[12] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3.
[13] Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 2020;367:1444–8.
[14] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181:271–80.
[15] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21.
[16] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet 1997;13:163.
[17] Wang Li, Wang Qinghua, Bai Heming, Liu Cong, Liu Wei, Zhang Yuanpeng, et al. EHR2Vec: representation learning of medical concepts from temporal patterns of clinical notes based on self-attention mechanism. Front Bioeng Biotechnol 2020;11:630.

[18] Ekins S, Bugrim A, Brovold L, Kirillov E, Nikolsky Y, Rakhmatulin E, et al. Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. Xenobiotica 2006;36:877–901.
[19] Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nat Rev Drug Discov 2020;19:149–50.
[20] Tong H, Faloutsos C, Pan JY. Random walk with restart: fast solutions and applications. Knowl Inf Syst 2008;14:327–46.
[21] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems 2006;1695:1–9.
[22] Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020;382:1708–20.
[23] Barton LM, Duval EJ, Stroberg E, Ghosh S, Mukhopadhyay S. COVID-19 autopsies, Oklahoma, USA. Am J Clin Pathol 2020;153:725–33.
[24] Gritti G, Raimondi F, Ripamonti D, Riva I, Landi F, Alborghetti L, et al. Use of siltuximab in patients with COVID-19 pneumonia requiring ventilatory support. medrxiv:20048561v4 [Preprint]. ?[cited ???]. Available from: https://www.medrxiv.org/content/10.1101/2020.04.01.20048561v4.full.pdf
[25] Xu X, Han M, Li T, Sun W, Wang D, Fu B, et al. Effective treatment of severe COVID-19 patients with tocilizumab. Proc Natl Acad Sci USA 2020;117:10970–5.
[26] Wu ZC, Wang X, Wei JC, Li BB, Shao DH, Li YM, et al. Antiviral activity of doxycycline against vesicular stomatitis virus in vitro. FEMS Microbiol Lett 2015;362:fnv195.
[27] Rothan HA, Mohamed Z, Paydar M, Rahman NA, Yusof R. Inhibitory effect of doxycycline against dengue virus replication in vitro. Arch Virol 2014;159:711–8.
[28] Sturtz FG. Antimurine retroviral effect of doxycycline. Methods Find Exp Clin Pharmacol 1998;20:643–7.
[29] Ren S, Guo LL, Yang J, Liu DS, Wang T, Chen L, et al. Doxycycline attenuates acrolein-induced mucin production, in part by inhibiting MMP-9. Eur J Pharmacol 2011;650:418–23.
[30] Shen L, Niu J, Wang C, Huang B, Wang W, Zhu N, et al. High-throughput screening and identification of potent broad-spectrum inhibitors of coronaviruses. J Virol 2019;93:e00023-19.
[31] Chan JF, Chan KH, Kao RY, To KK, Zheng BJ, Li CP, et al. Broad-spectrum antivirals for the emerging Middle East respiratory syndrome coronavirus. J Infect 2013;67:606–16.
[32] Shirey KA, Lai W, Scott AJ, Lipsky M, Mistry P, Pletneva LM, et al. The TLR4 antagonist Eritoran protects mice from lethal influenza infection. Nature 2013;497:498–502.
[33] Zheng YY, Ma YT, Zhang JY, Xie X. COVID-19 and the cardiovascular system. Nat Rev Cardiol 2020;17:259–60.
[34] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;395:565–74.
[35] Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. Lancet 2020;395:30–1. Corrected and republished from: Lancet 2020;395:1906.

Please cite this article as: Wang L et al., Novel approach by natural language processing for COVID-19 knowledge discovery, Biomedical Journal, https://doi.org/10.1016/j.bj.2022.03.011