**Prediction of Clinical Manifestations and Therapeutic Drugs of COVID-19 Using Natural Language Processing and Literature Data Mining**

Li Wang1,4\*, Lei Jiang2\*, Dongyan Pan3\*,Qinghua Wang1, Zijian Kang2,Haoran Tian2,Xuqiang Geng2,Jinsong Shao5,Wenjie Pan1,Jian Yin2,Yue Wang6,Weide Zhang7,Zhixiu Li8,Jun Zheng9,Wenxin Hu9,Yunbao Pan10,Dong Yu11, Shicheng Guo12,13, Wei Lu14#,Qiang Li15#,Huji Xu2#

1Medical School, Nantong University, Nantong, China

2Department of Rheumatology and Immunology, Shanghai Changsheng Hospital, Second Military Medical University, Shanghai, China

3Department of Ophthalmology, Shanghai Changhai Hospital, Second Military Medical University, Shanghai, China

4Research Center for Intelligence Information Technology, Nantong University, Nantong, China

5Publice Health School, Nantong University, Nantong, China

6Department of Histology & Embryology, Second Military Medical University, Shanghai, China

7Big Data and Artificial Intelligence Center, Zhongshan Hospital, Fudan University, Shanghai, China

8Translational Genomics Group, Institute of Health and Biomedical Innovation, Queensland University of Technology at Translational Research Institute, Princess Alexandra Hospital, Brisbane, Australia

9School of Data Science & Engineering, East China Normal University, Shanghai, China

10Department of Laboratory Medicine, Zhongnan Hospital of Wuhan University, Wuhan University, Wuhan, China

11Center for translational medicine, Second Military Medical University, Shanghai, China

12Department of Medical Genetics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53726, USA

13Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA

14NO.905 hospital Navy PLA, Shanghai, China

15Department of Respiratory and Critical Care Medicine, Shanghai East Hospital, Tongji University, Shanghai, China

\*Li Wang, Lei Jiang and Dongyan Pan contribute equally to this work.

#Corresponding author: Lu Wei, Li Qiang and Xu Huji contribute equally to this work.

Lu Wei, Email: smmuluwei@163.com

Li Qiang, Email: liqressh@hotmail.com

Xu Huji, Email: xuhuji@smmu.edu.cn

Abstract

2019 novel corona virus (2019-nCoV) has caused serious public health concerns. Although numerous basic research and clinical trials have been initiated in the past months, how the well estimated knowledge generated in previous practice to guide current study, clinical design and decision making is highly required. Herein, we developed a new high-throughput scientific research literature and data mining tool based on a novel natural language and artificial intelligence model MedE2vec. Utilizing our pipeline, we provided a well prediction on possible 2019-nCov related clinical manifestations and potential virus-targeted host organs. In addition, we generated an atlas of involved cell signaling and pathways related to 2019-nCOV invasive and infection. We also provided a prediction to the effectiveness of current available drugs in clinical practice. Our study provided valuable references for public health prevention, pathological mechanisms, clinical diagnosis and treatment and new drug development of new corona virus-related diseases. We suggest the MedE2vec-based AI tool utilizing artificial intelligence, natural language processing and literature data mining is of great significance in the face of the urgent task of research and design requirements for emerging pathogen prevention and control.

**Introduction**

Since December 2019, a special pneumonia first occurring in Wuhan, with a rapidly increasing number of patients, has quickly spread across China[1-3](#_ENREF_1" \o "Chen, 2020 #1326). A novel corona virus was immediately isolated from patient samples and identified as the 2019 novel corona virus (2019-nCoV)[4](#_ENREF_4" \o "Chen, 2020 #1311).The genomic analysis of 2019-nCoV showed 79.5% similarity to SARS-CoV and 96% similarity to bat-sourced CoV. 2019-nCoV belongs to the genus β corona virus. β corona virus is an encapsulated single-stranded RNA virus that can infect wildlife, herds, and human beings, resulting in occasional outbreaks and, more commonly, asymptomatic infections. The internal structure of the 2019-nCoV, the major protein, is chemically bound in a way similar to that of severe acute respiratory syndrome (SARS) or SARS corona virus[5](#_ENREF_5" \o "Lu, 2020 #1325). It was found by protein crystal structure analysis that 2019-nCoV S protein and SARS corona virus S protein had almost the same 3D structure in the Ras-binding domain (RBD domain). New studies have also confirmed that novel corona virus (SARS-CoV-2) entered cells via SARS-CoV receptor *ACE2* and that serine protease *TMPRSS*2 could activate the binding of S proteins and *ACE2* receptors. Comprehensive mechanisms of virus invasion and infection response are unclear.

In terms of clinical manifestations, 2019-nCoV patients will have fever, cough, retching, pneumonia, headache, and severe patients will have dyspnea, coagulation dysfunction and other symptoms[1](#_ENREF_1" \o "Chen, 2020 #1326)[6](#_ENREF_6" \o "Wang, 2020 #1279). However, a small number of patients have atypical symptoms, which bring difficulties for the clinical diagnosis. In the process of epidemic prevention and control of infectious disease with an extremely complex clinical manifestations, we discovered that the understanding, treatment measures and decision-making response to the burst disease were based on other events in the past and individual experiences of experts. This kind of decision-making method based on experiences has high efficiency and quick effect, but there are many problems, which will cause waste of social resources and some adverse consequences. For example, 250 clinical trials were carried out in a rush to meet the submission deadline, and there was no pharmacological and biological basis for the candidate treatment drugs. The obsolete candidate drugs were selected without considerations of the progress in the pharmaceutical field. In addition, novel corona virus pneumonia still lacks effective treatment. It is urgent to find effective drugs based on pathogenesis, clinical manifestations, organ involvement and past treatment experience. To solve these problems, we need a rapid and high-throughput analysis strategy based on the principle of evidence-based medicine, which can make full use of a large number of human research literature resources and public databases, and provide the support of literature theory or information data for the relevant research start-up and decision-making.

The establishment of big data analysis models for medical literature and public data relies on efficient and structured processing of human natural languages. The Bidirectional Encoder Representations from Transformers (BERT) model (reference here) of attention mechanism in natural language is an important data processing scheme reported in recent years. We used the analysis of special scenarios in the field of medicine to apply this mechanism evolution to vector generation tasks of medical biological entities (genes, diseases, drugs, etc.) and proposed MedE2vec models (reference here to explain what’s MedE2vec?). The model can transform medical biological entities in unstructured electronic medical records or medical literature into vectors. There are several medical biological entities (“Entity”), such as diagnosis, medication, symptoms, genes, molecules, and examination, in a patient's event (“Event”) of diagnosis and treatment, so the E of MedE2vec has two meanings: one is the patient's diagnosis and treatment event (“Event”), the other is the medical biological entity (“Entity”). The effectiveness of the study was shown in our previous research on lupus (papers about relevant working have been submitted. Here you can submit the manuscript to bioRxiv to cite it or else don’t mention it here).

Although COVID-19 is a new disease, as a group of human pathogens, we can still draw some clues from biological characteristics and similar infections such as SARS, MERS. For instance, the target of the new corona virus shared same protein encoded by *ACE2* and *TMPRSS2* to enter human cells[7](#_ENREF_7" \o "Zhou, 2020 #1401). Based on current medical knowledge, we could get candidate treatments for the new unknown diseases by doing data mining in all relevant research articles. In this study, based on natural language analysis and public medical database, with *ACE2*, *TMPRSS*2 as the main entrance, the AI analysis (Here AI should be explained how it was applied and what’s the relationship between AI and MedE2vec) of the following issues were performed using all available literatures: 1. possible clinical manifestations and involved target organs of COVID-19; 2. possible signaling pathways involved in COVID-19; 3. prediction of the effectiveness of current available drugs in clinical practice. This manuscript aimed to provide valuable references for public health prevention, pathological mechanisms, clinical diagnosis and treatment and new drug development of new corona virus-related diseases. (Introduction is too long, later we need to short it at least by 50%)

**Methods**

**Data collection,**

*ACE2* and *TMPRSS*2 were used as keywords to conduct literature retrieval in PubMed public database, respectively. All the title and abstracts were obtained with our customed script (share the script in github). Medical biological related entities in abstracts were extracted with MetaMap[8](#_ENREF_8" \o "Aronson, 2010 #1403) while entity relationship was analyzed with SemRep[9](#_ENREF_9" \o "Rindflesch, 2003 #1402). MedE2vec (reference for MedE2vec) was applied to calculate the entity correlation. MedE2vec is a modified Med2vec method relying on attention mechanism to extract features. MedE2vec 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. MedE2vec used the Adadelta optimizer (reference here). The number of attention heads in the self-attention mechanism was eight, and the number of vector dimensions was 512. MedE2vec was trained 20 epochs for the best result (Why to set 20 should be explained).

**Analysis and validation of clinical manifestations**

Keywords including SARS-CoV-2, 2019-nCoV, COVID-19, SARS-CoV, SARS, MARS-CoV, and MARS were used to conduct literature retrieval in PubMed and CNKI (details or share scripts). Two doctors with rich clinical experience selected literatures concerning clinical manifestations and organ involvement, extracted clinical manifestations and organ damage, and calculated the incidence according to the number of cases to verify with the result of computer analysis (what kind of computer analysis? Show details or show it more explicitly) based on natural language.

**Bioinformatics analysis and drug correlation prediction**

Molecules and protein tokens in the top 200 entities (why top 200? Not 500?) were extracted for enrichment analysis of signaling pathways using direct interaction algorithm of MetaCoreTM (Clarivate Analytics) (p<0.05, here what’s P it is? What kind of test was used to obtain this P-value?). According to the results, pathways were verified in the 21 reported drugs based on host-based treatment strategies using MetaCore database[10](#_ENREF_10" \o "Li, 2020 #1414). Besides, top 50 pathway (why top 50?) related MetaCore collected drugs were summarized to support further clinical trials.

Statistical analysis

Describe all the P-values here and which software were used to conduct these tests, R or SPSS or what?

**Results**

**Vector Construction and Optimization of Entity Correlation Calculation**

We received a total of 1,025 papers related to *TMPRSS*2 and 1,912 papers related to *ACE2 (upload and share these 2937 abstract in github).* Entities and relationships analysis to abstract were formed respectively with the processing by MetaMap and SepRep. Finally, vector calculation and sequencing were carried out, and candidate analytical entities were labeled by human body, organ, disease, protein, gene, organic/inorganic molecules and drug. We notice the top 200 candidate entities provided useful information for the clinical research and clinical trials (supplementary table to show top 500 entitles), such as xxx, xxx, xxx, xxx. [We need P-values for each candidate here]. Clinical manifestations identified by the MedE2vec showed xxxx indicating xxxx

**Combined data analysis of natural language and bioinformatics based on *ACE2* and *TMPRSS*2**

Firstly, we processed the top 200 related entities obtained respectively by setting *ACE2* and *TMPRSS*2 as the key words, which belonged to 210 categories of meta map token. According to the research needs, 40 tokens related to human body, organs, diseases, proteins, genes, organic and inorganic molecules and drugs were selected and classified into four categories: molecules, proteins, organs/systems and diseases. A total of 27 related molecular/protein entities, 33 organ and system entities and 44 disease entities were obtained by *ACE2*, while 43 related molecular/protein entities, 17 organ and system entities, and 32 disease entities by *TMPRSS*2. Afterwards, analysis of clinical manifestations and bioinformatics were performed respectively (**Figure 1**).

In this study, 33 *ACE2*-related organs and systems were obtained, including intestine, renin-angiotensin system, renin-angiotensin-aldosterone system, liver, skeleton, skeletal muscle, respiratory organs, brain, autonomic nervous system, kidney, limbs, cardiovascular system, and pancreas. The top five were listed as follows: renin-angiotensin system had the highest level of 0.85, followed by heart and kidney both at 0.68; spine had a correlation at 0.55; respiratory system had a correlation at 0.42, similar to brain at 0.41 (**Table 1**). Meanwhile, 17 organs and systems relevant to *TMPRSS*2 were obtained, including the urinary system, nervous system, respiratory system, bone and limbs, oral cavity, endocrine system, etc. The top five were listed as follows: prostate, with an OR-value at 0.83, hippocampus proper at 0.39, lung at 0.21, urethra and skeleton at 0.20 (**Table 2**). Forty-four *ACE2*-related disease entities were obtained, including several types of diseases affecting cardiovascular, kidney, endocrine, brain, eye, and so on, of which macrophage activation syndrome had the highest correlation at 0.77, followed by SARS at 0.72, hypertensive disease at 0.67, Severe Acute Respiratory Syndrome at 0.63, and cardiac arrest at 0.56. (**Table 3**)

The diseases were classified by the affected organs and systems, and the average correlation of respiratory system related diseases was 0.34. According to the correlations from high to low, the sequence was that of SARS, Adult Respiratory Distress Syndrome, pulmonary fibrosis and pulmonary edema. The average correlation of cardiovascular-related diseases was 0.33, and the sequence was that of high blood pressure, cardiac arrest, cardiac fibrosis, heart failure, etc. from high to low. Eye-related diseases had an average correlation at 0.29, including fish-eye disease and uveitis. The average correlation of kidney-related diseases was 0.28, including chronic kidney disease, diabetic nephropathy, renal fibrosis, and nephrotic syndrome. The average correlation of endocrine-related diseases was 0.25, and the sequence was that of diabetes, obesity and hypernatemia from high to low. The average correlation of brain-related diseases was 0.23, including Alzheimer's disease, ischemic stroke, and cerebrovascular disease.

Thirty-two *TMPRSS*2-related diseases entities were obtained, including several types of diseases affecting cardiovascular, autoimmune and other systems, of which cardiac arrest had the highest correlation at 0.41, followed by flu at 0.37, infectious disease at 0.36, prostatic disease at 0.36, and hypersensitivity at 0.31. We found that among the diseases predicted above, the number of infectious diseases was the highest, with an average OR of 0.29, followed by autoimmune diseases including xerophthalmia, autoimmune response, hypersensitivity with an average OR of 0.22; urinary system diseases had an average OR of 0.22, with the prostate mainly affected; endocrine diseases mainly included diabetes, with an OR of 0.19; and respiratory system diseases mainly included SARS, with an OR of 0.19. (**Table 4**). Meanwhile, we found that the gene was highly associated with virus infections, including Orthomyxoviridae, SARS corona virus, MERS corona virus, Influenza A virus, Metapneumovirus, etc. (**Table 5**)

In order to test the reliability of AI analysis, we carried out literature search in PubMed, CNKI, and other databases. A total of 34 references and 11817 cases related to clinical manifestations and organ involvement were selected, in which 2019-nCOV (4 references, 1278 cases), SARS (17 references, 7671 cases), and MERS (13 references, 2868 cases) were included. In addition, there were 24 references and 64 cases of autopsy or tissue and organ reports for COVID-19 and SARS.

According to the statistics, respiratory system involvement accounted for 42.8%, 39.5% and 64.1% of the three corona virus-related diseases, respectively; digestive system involvement accounted for 3.5%, 13% and 20.5%; eye involvement was reported to account for 4.6% in 2019-nCOV. Fever symptoms accounted for 27.9%, 37.4% and 31.1% respectively. In addition, there were headache, myalgia, fatigue, chill, vertigo and other symptoms. (**Table 6**)

We found that respiratory system involvement was the most common, which was also confirmed in our algorithm (*ACE2*, OR=0.34; *TMPRSS*2, OR=0.19). The algorithm also indicated the symptoms of eye involvement (*ACE2*, OR=0.29), and these clinical symptoms also had higher frequency in patients with 2019-COV (4.6%). The algorithm also found that the brain (0.23), the endocrine system (*ACE2*, OR=0.25; *TMPRSS*2, OR=0.19) and urinary system (*ACE2*, OR=0.28; *TMPRSS*2, OR=0.22) might have a certain degree of involvement.

**Bioinformatics analysis and drug prediction.**

Through the literature mining of natural language and public gene databases, we screened 27 molecules proteins based on ACE II keyword, and found that acetylcholinesterase, vasoactive intestinal peptide receptor

SLC33A1, CTSC, apalein peptide, AGTR1, angiotensin Ⅱtype I receptor, MAPK3, IL-6, etc. were highly correlated (Table 7). Moreover, the screening based on *TMPRSS*2 keyword discovered 43 molecular/protein related entities, among which PTEN, NAA50, ERG, PSAT1, GGH, etc. had higher correlations. (**Table 8**)

A total of 745 pathways were obtained by enriching 27 related molecules/proteins of ACE II. The pathways of TOP50 were listed in the table, including Oxidative stress\_ROS-induced cellular signaling, Angiotensin II/AGTR1 signaling via p38, ERK and PI3K, IL-17-induced mucin expression in CF airways, etc. (**Figure 2**) Forty-three related molecular/protein pathways of *TMPRSS*2 were enriched, and a total of 686 pathways were obtained. The pathways of TOP50 were listed in the table, including Cell cycle\_Role of 14-3-3 proteins in cell cycle regulation, Immune response\_HMGB1 release from the cell, Signal transduction\_PTEN pathway, etc. (**Figure 3**)

In order to further verify the significance of our prediction genes and pathways based on natural language for drugs related to 2019-nCOV virus therapy, we used MetaCore databases to search for *ACE2* and *TMPRSS* as keywords to find out the top 50 pathways with the most significant correlation. We further verified the 21 potential therapeutic drugs of the nCOV pneumonia based on Host-based treatment strategies reported in the literature. The pathways involved in the 21 drugs were searched by MetaCore databases. Among them, 10 drugs had the research data of related pathways, and a total of 623 pathways were obtained. Among them, three kinds of drug-related pathways exist in the *ACE2* and *TMPRSS*2-related pathways obtained by our calculation and analysis, which are interferon, rapamycin and K22 respectively. The specific results were shown in **Table 9**.

Previous studies indicated that *ACE2* and *TMPRSS* were important molecules in the process of 2019-nCOV virus infection. Therefore, *ACE2* and *TMPRSS* genes were used to find out the associated clue genes in order to obtain the enrichment pathway and find the treatment. We found 21 kinds of therapeutic drugs for 2019-nCOV by searching the literature. Once again, the pathways involved in these 21 drugs were searched by the MetaCore databases. Among them, 10 drugs had related pathways, and a total of 623 pathways were obtained. After comparing the gene-related pathways with drug-related pathways and removing the disease pathways which were not related to virus infection (such as tumors), it was found that there were 11 *ACE2*-related pathways and 9 *TMPRSS*-related pathways, mainly interferon, rapamycin and K22.

With the application of MetaCore databases with *ACE2* and *TMPRSS* set as keywords, the top 50 pathways with the most significant correlation were found (p < 0.05). There were altogether 427 drugs found in the top 50 pathways (**Table 10**). In terms of drug categories, they were all respiratory, immune related, infectious or inflammatory drugs. For example, Beclabuvir and Saquinavir, can bind to MPRO protein, a key enzyme involved in the extensive proteolytic processing of the virus’ polyproteins. Azelastine, which has mast cell-stabilizing properties that prevent the release of interleukin-6, tryptase, histamine, and TNF-alpha from mast cells, is an antihistamine available for the treatment of allergic and vasomotor rhinitis. Pictilisib, a potent pan inhibitor of class I catalytic subunits of PI3K, is available for cancer research.

**Discussion**

Since the beginning of the 21st century, many emerging pathogens have caused major public health disease outbreaks, such as SARS, avian influenza, MARS, Ebola virus, and Cov2019, have brought great threat to social and economic development and human life safety. However, there is an extremely urgent need for the research on the pathogenic mechanism, prevention strategy, diagnosis and treatment of pathogens in each outbreak, and the scientific research work promoted by the traditional exploration-hypothesis is often difficult to keep up with the urgent need for epidemic prevention and control. On the other hand, there are a large number of scientific research literatures and data in the past, and if the valuable scientific research data are mined in an orderly manner, research ideas and hypotheses may be formed very quickly to guide the next step of research work. Therefore, the establishment of a new high-throughput scientific research literature and data mining tool based on natural language is of great significance in the face of the urgent task of research and design requirements, such as novel corona virus prevention and control.

In the strategy of natural language exploration, we adopted vector analysis strategy driven by keyword, in which keyword was the important start of the whole process of vector analysis. In this study, ACE II and TMPRESS, the key invasion target proteins of new corona virus, were selected as the main key entries, and the main findings of this paper were obtained. This system analysis strategy driven by keyword was very suitable for natural language mining in which research topics switched fast. We could further track the new research targets according to the new research reports, or quickly explore the results in other types of pathogens, public health events and other social events, and get the systematic analysis results within a few hours, which were significant technical advantages and features of this strategy.

In the reality field of medicine, this study analyzed the key research needs related to the prevention and control of new corona virus outbreak. First of all, the possible invasion manner of virus and the analysis of the involved organs were closely related to the prevention and control of virus as well as the potential clinical manifestations, which was helpful for the clinical diagnosis and detection of disease changes. By analyzing the clinical manifestations of the three diseases, we found that the most common symptoms of the patients were fever and respiratory system, such as dry cough, nasal obstruction, runny nose, sore throat, dyspnea, etc. (Respiratory system symptoms accounted for about 45.7%), which was highly consistent with the respiratory system involvement predicted by our algorithm analysis with *ACE2* and TEMPRESS2 as the key words (*ACE2* OR= 0.34, TEMPRESS2 = 0.19); secondly, patients often suffered from nausea, vomiting, diarrhea, anorexia and other digestive system symptoms (Symptoms accounted for about 13.3%), which was also accurately predicted by our algorithm (*ACE2* OR= 0.24). At the same time, the algorithm also suggested that the possible symptoms of eyes (*ACE2* OR= 0.29, TEMPRESS2 = 0.19), and this clinical symptom also had a higher frequency in 2019-CoVID patients (Symptoms accounted for 4.6%). Symptoms of the cardiovascular system (*ACE2* OR=0.33), kidney (*ACE2* OR=0.28), endocrine system (*ACE2* OR=0.25,TEMPRESS2 OR=0.19) and brain (*ACE2* OR=0.23) were also found in the algorithm prediction, and the involvement of these systems was mentioned in the autopsy or histopathological case reports of COVID-19 and SARS, which showed that our algorithm had a reliable prediction performance. Further results need to be confirmed by collecting more clinical data.

Our results showed that novel corona virus pneumonia was associated with *ACE2* in adults, and the correlation level was 0.3. Adult respiratory distress syndrome was characterized by refractory hypoxemia and high mortality. It has been reported in patients with novel corona virus pneumonia, with an incidence of 3.4%. [11](#_ENREF_11" \o "Guan, 2020 #1412)Moreover, we also found that hypertension and diabetes were also associated with this disease, which was consistent with the report that novel corona virus pneumonia had a high occurrence in patients with hypertension and diabetes. Therefore, our prediction model can provide reference for clinical diagnosis and disease development.

The second direction was the pathogenesis of virus, especially the molecular mechanism. We found that novel corona virus pneumonia was directly associated with 70 molecules/proteins associated with *ACE2* and *TMPRSS*2, including IL-6 (OR=0.26), and an increasing level of IL-6 has been detected in the blood of new corona virus pneumonia patients, which may be involved in the new corona virus related inflammatory reaction.

Through the enrichment of related molecular pathways of ACEII and *TMPRSS*2 protein, we found that the TOP50 pathway contained signal transduction, that is, PTEN pathway. PTEN antagonized Phosphoinositide-3-kinase (PI3K), dephosphorylating PtdIns(3,4,5)P3 and inhibited V-akt murine thymoma viral oncogene homolog 1 (PI3K)/AKT-mediated signaling pathways. [12](#_ENREF_12" \o "Dahia, 2000 #1404)As a result, PTEN prevented IGF-1 signaling[13](#_ENREF_13" \o "Knobbe, 2002 #1405) [14](#_ENREF_14" \o "Lackey, 2007 #1406) and activation of the EGFR-induced mTOR[15](#_ENREF_15" \o "Mellinghoff, 2007 #1408). PI3K Akt signaling pathway is involved in the invasion process of virus to host cells.[16](#_ENREF_16" \o "Ehrhardt, 2006 #1409) The regulation of PI3K Akt signaling pathway is related to viral uptake, and PI3K Akt signaling pathway may be necessary for viral endocytosis.[17](#_ENREF_17" \o "Izmailyan, 2012 #1410) In acute infection, some viruses promote the short-term survival of infected cells through activation of PI3K Akt signaling pathway, which is conducive to the effective propagation of viruses before cell death.[18](#_ENREF_18" \o "Lindemans, 2006 #1411) Therefore, PTEN inhibition of PI3K/Akt and downstream mTOR may be helpful to prevent virus from invading host cells. On February 14, Demiurge Inc. announced that PI3K inhibitor was expected to become a new type of effective and safe candidate therapeutic drug for COVID-19, which was based on the epidemiological, clinical and biological characteristics of COVID-19, by establishing the cross disease model in the field of oncology and virology. Although we used different methods, we still find the potential significance of PTEN path, which proved the reliability of our model from another side.

We also found IL-17-induced mucin expression in CF airways in TOP50 pathway. Inflammatory cytokine IL-17, together with IL-6 and IL-1 beta, can enhance expression of mucin genes in CF airways, contributing to recurring cycles of infection followed by increased expression of mucins that culminates in airway obstruction with mucus. Although this pathway was found in Cystic Fibrosis (CF) patients, we still found that there were a lot of mucus in the alveoli in the autopsy of new corona virus pneumonia patients. Is there any excessive expression and secretion of mucin after inflammatory factors stimulation? It requires us to carry out further researches. Therefore, the discovery of these pathways can provide important candidate research targets for basic researches.

The third direction was host-based treatment strategies. Because the new drug clinical experiment requires strict approval process and validation cycle, it is difficult to apply it immediately in the epidemic period. Based on the analysis of related pathways, we screened drugs according to TOP50 pathway, suggesting that a batch of approved drugs could play a role in virus prevention and treatment of complications, and these results will greatly shorten the clinical application of drugs The application cycle may provide important reference data for the control of novel corona virus. Rapamycin, the specific mTOR inhibitor, was also included in the drugs we predicted through pathway, which also proved the effectiveness of our algorithm from another side. In the future, we will analyze more direction needs and obtain more reference data, which all reflect the potential application value of our research strategy.

Of course, there are still some defects in the strategy of natural language document discovery. For example, the current openness of public literature was limited, and most of our data mining was based on abstracts, which leads to the result that analysis of data were only base on the core data discovery of literature, there was still a certain possibility of incomplete mining. In addition, our mining strategy was mainly aimed at natural language, so it was greatly influenced by the subjective judgment of the conclusion of the author of the literature. In the future, it is possible for us to have better judgment if we weight the literature with different sources and different reliability. Therefore, it may be a better strategy to combine the discovery of natural language based on literature and the discovery of bioinformatics data from some public databases to repeatedly verify the conclusions. We have made a part of attempts in this study, and we will develop the above strategy joint operation tools and platforms in more studies, so as to improve the reliability and validity of related analysis.

**Acknowledgment**

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

**REFERENCE**

1. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020.

2. Chang, Lin M, Wei L, et al. Epidemiologic and Clinical Characteristics of Novel Coronavirus Infections Involving 13 Patients Outside Wuhan, China. Jama 2020.

3. Jiang S, Xia S, Ying T, et al. A novel coronavirus (2019-nCoV) causing pneumonia-associated respiratory syndrome. Cellular & molecular immunology 2020.

4. Chen L, Liu W, Zhang Q, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerging microbes & infections 2020;9(1):313-19.

5. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020.

6. Wang D, Hu B, Hu C, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. Jama 2020.

7. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020.

8. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA 2010;17(3):229-36.

9. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics 2003;36(6):462-77.

10. Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nature reviews Drug discovery 2020;19(3):149-50.

11. Guan WJ, Ni ZY, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. The New England journal of medicine 2020.

12. Dahia PL. PTEN, a unique tumor suppressor gene. Endocrine-related cancer 2000;7(2):115-29.

13. Knobbe CB, Merlo A, Reifenberger G. Pten signaling in gliomas. Neuro-oncology 2002;4(3):196-211.

14. Lackey J, Barnett J, Davidson L, et al. Loss of PTEN selectively desensitizes upstream IGF1 and insulin signaling. Oncogene 2007;26(50):7132-42.

15. Mellinghoff IK, Cloughesy TF, Mischel PS. PTEN-mediated resistance to epidermal growth factor receptor kinase inhibitors. Clinical cancer research : an official journal of the American Association for Cancer Research 2007;13(2 Pt 1):378-81.

16. Ehrhardt C, Marjuki H, Wolff T, et al. Bivalent role of the phosphatidylinositol-3-kinase (PI3K) during influenza virus infection and host cell defence. Cellular microbiology 2006;8(8):1336-48.

17. Izmailyan R, Hsao JC, Chung CS, et al. Integrin beta1 mediates vaccinia virus entry through activation of PI3K/Akt signaling. Journal of virology 2012;86(12):6677-87.

18. Lindemans CA, Coffer PJ, Schellens IM, et al. Respiratory syncytial virus inhibits granulocyte apoptosis through a phosphatidylinositol 3-kinase and NF-kappaB-dependent mechanism. Journal of immunology 2006;176(9):5529-37.

**Figure legends**

Figure 1. Workflow of combined data analysis of natural language and bioinformatics based on *ACE2* and *TMPRSS*2 as key words

Figure 2. TOP 50 pathways of ACE II

Figure 3. TOP 50 pathways of *TMPRSS*2

Table 1 *ACE2*-related organs and systems

Table 2 *TMPRSS*2-related organs and systems

Table 3 *ACE2*-related diseases

Table 4 *TMPRSS*2-related diseases

Table 5 *TMPRSS*2-related viruses

Table 6 Literatures on clinical manifestations and organ involvement of 2019-nCOV, SARS and MERS

Table 7 ACE II -related molecules / proteins

Table 8 *TMPRSS*2 -related molecules / proteins

Table 9 Verification of the 21 reported potential therapeutic drugs

Table 10 Potential therapeutic drugs found in the top 50 pathways